RICE UNIVERSITY

Moment Matching and Modal Truncation for Linear Systems

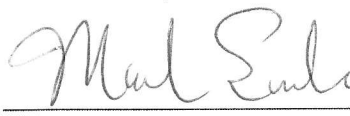
by

A.J. Hergenroeder

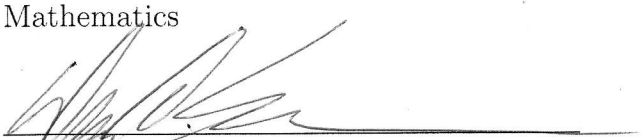
A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Master of Arts

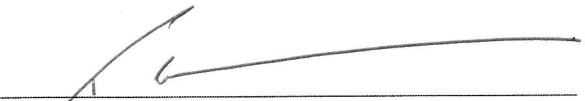
Approved, Thesis Committee:



Mark Embree, Chair
Professor of Computational and Applied Mathematics



Dan Sorensen
Noah G. Harding Professor of
Computational and Applied Mathematics



Tim Warburton
Associate Professor of Computational and Applied Mathematics

HOUSTON, TEXAS

November, 2012

ABSTRACT

Moment Matching and Modal Truncation for Linear Systems

by

A.J. Hergenroeder

While moment matching can effectively reduce the dimension of a linear, time-invariant system, it can simultaneously fail to improve the stable time-step for the forward Euler scheme.

In the context of a semi-discrete heat equation with spatially smooth forcing, the high frequency modes are virtually insignificant. Eliminating such modes dramatically improves the stable time-step without sacrificing output accuracy. This is accomplished by modal filtration, whose computational cost is relatively palatable when applied following an initial reduction stage by moment matching. A bound on the norm of the difference between the transfer functions of the moment-matched system and its modally-filtered counterpart yields an intelligent choice for the mode of truncation.

The dual-stage algorithm disappoints in the context of highly nonnormal semi-discrete convection-diffusion equations. There, moment matching can be ineffective in dimension reduction, precluding a cost-effective modal filtering step.

Acknowledgements

I thank God for life, for this opportunity to study applied mathematics, and for the beautiful theory contained in this thesis.

I thank my thesis advisor, Dr. Mark Embree, whose mathematical brilliance is matched by, among other qualities, his contagious humility and his passion for his students. Over the years, he has continually shown genuine interest both in my mathematical and my personal development. I thank Dr. Embree for the patience, concern, and respect that he has shown me. I am one of the luckiest students at Rice University to have worked with him.

I thank Rice University and its faculty for providing me with this extraordinary experience.

I thank Dr. Tim Warburton for his thoughtful input on my research, for serving on my thesis committee, for several conversations regarding PDEs and ODEs, and for offering me great insight into industrial mathematics.

I thank Dr. Danny Sorensen for serving on my thesis committee, for his helpful comments on my thesis, and for his fantastic course in numerical linear algebra.

I thank Dr. Russell Carden, one of the most intelligent and humble people that I have met. Russell has generously offered me valuable input and encouragement in my research. I thank Russell for this and for being a great friend.

I thank Dr. Joanna Papakonstantinou. It was through her guidance that I unlocked my passion and skill for mathematics in high school. Joanna subsequently helped to recruit me to the CAAM Department. Without her, none of this research could have happened. I thank Joanna for being one of the best friends and mentors that I have ever had.

I thank Dr. Richard Tapia, whose creative sense of humor was refreshing. I thank him for his excellent Optimization Theory course, which exposed me to a portion of applied mathematics outside of my research, and for sharing many of his life experiences with me.

I thank Drs. Laurie Heyer, Ben Klein, David R. Larson, Donna Molinek, and John Swallow, each of whom provided invaluable mentoring to me as an undergraduate and encouraged me to pursue graduate study.

I thank Dr. Bob Hardt for a fascinating course in partial differential equations. He taught me a great deal, as evidenced in Chapters 2 and 3. I thank him for his friendship and patience.

I thank Drs. David Damanik, Frank Jones and Jim Tour for their friendships and mentoring during my time at Rice.

I thank Dr. Jennifer Young for her mentoring of me as CAAM 210's Head Lab TA. Jen's door was always open to me, and her warm nature made her so approachable.

I thank Dr. Bill Symes, for his warm attitude and for offering me much guidance in analysis during my first year at Rice.

I thank my friends Jeff Hokanson and Charles Puelz for helping me practice for my thesis defense. I also thank some of my friends in the sciences, Dr. Ricardo Alonso, Dr. Harbir Antil, Reid Atcheson, Dr. Thomas Callaghan, Jorge Castañón, Phillip Compeau, Wei Deng, Bosen Du, Yin Huang, Dr. Drew Kouri, Mark Lai, Dr. Rami Nammour, Nabor Reyna, Shirin Sardar, Adam Topaz, Toni Tullius, Xin Wang, Yingpei Wang, Brant West, Meagan Whaley, and Xin Yang, each of whom has been of great help to me at some point during my academic career.

I thank Dr. Steve Cox for his dedication as the Principal Investigator of the NSF VIGRE grant that funded me. I also thank the NSF, the CAAM Department, and Rice University for providing me with such generous funding.

I thank Daria Lawrence for providing guidance on completing this degree. I thank Brenda Aune, Jennifer Trevino, and Ivy Gonzalez for providing great support to me.

I thank Dad, Mom, Alicia, Georgene and Grandma for loving, supporting, and encouraging me all of my life. I love you.

Contents

To Dad, Mom, Alicia, Georgene, and Grandma,
and to Him Whose life, death, and resurrection give me hope.

Chapter 1

Reduced-Order Models

1.1 Motivating Dimension Reduction

Consider the linear dynamical system

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) \quad \text{for } t \geq 0, \\ \mathbf{x}(0) &= \mathbf{x}_0, \end{aligned}$$

where $\mathbf{x} : \mathbb{R} \to \mathbb{C}^n$ is zero at negative times, and $\mathbf{A} \in \mathbb{C}^{n \times n}$. The system has the exact solution

$$\mathbf{x}(t) = e^{t\mathbf{A}}\mathbf{x}_0$$

for $t \geq 0$, which can be approximated using a numerical integration rule, such as the forward Euler scheme. Taking a fixed time-step $h > 0$ with corresponding times $t_j := jh$ ($j \in \mathbb{N}$), forward Euler generates $\mathbf{x}_j \approx \mathbf{x}(t_j)$ using the scheme

$$\mathbf{x}_{j+1} := \mathbf{x}_j + h\mathbf{A}\mathbf{x}_j. \tag{1.1}$$

Suppose that all eigenvalues of $\mathbf{A}$ have negative real part, so that $\lim_{t\to\infty} \mathbf{x}(t) = 0$ for

all possible choices of $\mathbf{x}_0$. To ensure that $\lim_{j\to\infty} \mathbf{x}_j = \mathbf{0}$, $|h\lambda + 1| < 1$ must be true for all $\lambda \in \sigma(\mathbf{A})$. This stability requirement often becomes burdensome as n grows.

Suppose, for example, that $\mathbf{A} \in \mathbb{R}^{n\times n}$ is the discrete Laplacian formed through the second order finite difference discretization. Then $\mathbf{A}$ satisfies $\sigma(\mathbf{A}) \subset (-\infty, 0)$, and thus forward Euler's stability requirement amounts to

$$h < \frac{2}{\rho(\mathbf{A})},$$

where $\rho(\mathbf{A}) := \max_{\lambda\in\sigma(\mathbf{A})} |\lambda|$ is the *spectral radius* of $\mathbf{A}$. Yet $\rho(\mathbf{A}) = O(n^2)$ (see (2.30)), making stable integration impractical for large n.

However, it is uncommon for the highest frequency eigenvectors of $\mathbf{A}$, i.e., those corresponding to the largest magnitude eigenvalues of $\mathbf{A}$, to contribute significantly to $\mathbf{x} : \mathbb{R} \to \mathbb{R}^n$, in particular when $\mathbf{x}_0$ is the discretization of a smooth function. Here lies a key motivation for reduced-order models. In such cases, one may eliminate the least relevant modes of $\mathbf{A}$ — those that make the time-step restrictions difficult — while making minimal sacrifice in accuracy. One can dramatically increase the stable time-step h for the forward Euler scheme in doing so.

This thesis focuses heavily upon improving the stable time-step for the forward Euler scheme. Yet the results hold for the more general setting of any explicit scheme whose absolute stability region has a non-trivial intersection with the negative real axis.

Much of the notation used in Chapter 1 originates in [1] and [7].

1.2 Moment Matching

Now consider the linear, time-invariant (LTI) system

$$\Sigma := \left(\begin{array}{lcll} \dot{\mathbf{x}} & = & \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} & \text{for } t \geq 0 \\ \mathbf{y} & = & \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u} & \text{for } t \geq 0 \\ \mathbf{x}(0) & = & \mathbf{x}_0 & \end{array} \right), \tag{1.2}$$

where

$$\begin{aligned} \mathbf{A} &\in \mathbb{C}^{n \times n}, \\ \mathbf{B} &\in \mathbb{C}^{n \times m}, \\ \mathbf{C} &\in \mathbb{C}^{p \times n}, \\ \mathbf{D} &\in \mathbb{C}^{p \times m}, \end{aligned}$$

and where $\mathbf{x}$ lies in the *state space*

$$\mathbb{X} := \left\{ \mathbf{x} : \mathbb{R} \to \mathbb{C}^n, \mathbf{x}|_{(-\infty,0)} = \mathbf{0} \right\},$$

$\mathbf{u}$ lies in the *input space*

$$\mathbb{U} := \left\{ \mathbf{u} : \mathbb{R} \to \mathbb{C}^m, \mathbf{u}|_{(-\infty,0)} = \mathbf{0} \right\},$$

and $\mathbf{y}$ lies in the *output space*

$$\mathbb{Y} \;:=\; \left\{\mathbf{y} : \mathbb{R} \to \mathbb{C}^p, \mathbf{y}|_{(-\infty,0)} = \mathbf{0}\right\}.$$

This system Σ is more compactly represented using the notation

$$\Sigma := \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}$$

employed in [1].

Moment matching model reduction involves finding a system $\widehat{\Sigma}$

$$\widehat{\Sigma} := \begin{pmatrix} \widehat{\mathbf{A}} & \widehat{\mathbf{B}} \\ \widehat{\mathbf{C}} & \mathbf{D} \end{pmatrix}$$

with state dimension $\widehat{n} \ll n$, where

$$\begin{aligned} \widehat{\mathbf{A}} &\in \mathbb{C}^{\widehat{n}\times\widehat{n}}, \\ \widehat{\mathbf{B}} &\in \mathbb{C}^{\widehat{n}\times m}, \\ \widehat{\mathbf{C}} &\in \mathbb{C}^{p\times\widehat{n}}, \\ \mathbf{D} &\in \mathbb{C}^{p\times m}, \end{aligned}$$

which generates output $\widehat{\mathbf{y}} \approx \mathbf{y}$ when common input $\mathbf{u} \in \mathbb{U}$ is used to drive both Σ and $\widehat{\Sigma}$.

1.2.1 The Laplace Transform of the System Output

For any $f \in \mathrm{L}^2(\mathbb{R})$, the *Laplace transform* at point $s \in \mathbb{C}$ is

$$(\mathcal{L}f)(s) := \int_0^\infty f(t)e^{-st}dt,$$

and the *Fourier transform* at point $\omega \in \mathbb{C}$ is

$$\begin{aligned}(\mathcal{F}f)(\omega) &:= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} f(\xi)e^{-i\xi\omega}d\xi \\ &= \left\langle f(\xi), \frac{e^{i\xi\omega}}{\sqrt{2\pi}}\right\rangle,\end{aligned}$$

where the L^2 inner product $\langle\cdot,\cdot\rangle : \mathrm{L}^2(\mathbb{R}) \times \mathrm{L}^2(\mathbb{R}) \to \mathbb{R}$ is defined

$$\langle f, g\rangle := \int_{\mathbb{R}} f(x)\overline{g(x)}dx.$$

Observe that for any output $\mathbf{y} \in \mathbb{Y} \cap \mathrm{L}^2(\mathbb{R})$,

$$\begin{aligned}(\mathcal{L}\mathbf{y})(i\omega) &= \int_0^\infty \mathbf{y}(t)e^{-i\omega t}dt \\ &= \int_{\mathbb{R}} \mathbf{y}(t)e^{-i\omega t}dt \\ &= \sqrt{2\pi}(\mathcal{F}\mathbf{y})(\omega),\end{aligned}$$

relating the Laplace and Fourier transforms of $\mathbf{y}$.

Now the set of functions $\left\{\frac{e^{i\omega x}}{\sqrt{2\pi}} : \mathbb{R} \to \mathbb{C}\right\}_{\omega\in\mathbb{R}}$ can be used to represent any func-

tion $f \in \mathrm{L}^1(\mathbb{R})$ that has only bounded discontinuities and whose Fourier transform $(\mathcal{F}f)(\omega)$ exists for all $\omega \in \mathbb{R}$, in the sense that for all $t \in \mathbb{R}$,

$$\begin{aligned} \frac{1}{2}\left(\lim_{r\downarrow t} f(r) + \lim_{r\uparrow t} f(r)\right) &= \int_{\mathbb{R}} (\mathcal{F}f)(\omega)\frac{e^{i\omega t}}{\sqrt{2\pi}}d\omega \\ &= \int_{\mathbb{R}} \left\langle f(s), \frac{e^{i\omega s}}{\sqrt{2\pi}} \right\rangle_{s\in\mathbb{R}} \frac{e^{i\omega t}}{\sqrt{2\pi}}d\omega \end{aligned}$$

(see, e.g., [3, p. 9-10], where the result follows from a change of variables on the double integral at the bottom of p. 9). Namely, when f is continuous at x, the left hand side of the previous equation simplifies to $f(t)$. Thus for $\mathbf{y} \in \mathbb{Y}$ satisfying the preceding assumptions, for all $t \in \mathbb{R}$ where $\mathbf{y}$ is continuous, $\mathbf{y}(t)$ can be expressed using its Laplace transform by

$$\begin{aligned} \mathbf{y}(t) &= \int_{\mathbb{R}} (\mathcal{F}\mathbf{y})(\omega)\frac{e^{i\omega t}}{\sqrt{2\pi}}d\omega \\ &= \frac{1}{2\pi}\int_{\mathbb{R}} (\mathcal{L}\mathbf{y})(i\omega)e^{i\omega t}d\omega. \end{aligned}$$

Hence it is reasonable to expect that $\widehat{\mathbf{y}} \approx \mathbf{y}$ when $(\mathcal{L}\widehat{\mathbf{y}})(s) \approx (\mathcal{L}\mathbf{y})(s)$ at those frequencies $s \in i\mathbb{R}$ that most significantly contribute to the output $\mathbf{y}$ — those where $\|(\mathcal{L}\mathbf{y})(s)\|$ is largest. Model reduction by moment matching is based on this idea.

1.2.2 The Transfer Function $\mathcal{H} : \mathbb{R} \to \mathbb{C}^{p\times m}$ and its Moments Near 0

For the system

$$\Sigma := \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix},$$

it can be shown that

$$(\mathcal{L}\mathbf{y})(s) \;=\; \mathcal{H}(s)(\mathcal{L}\mathbf{u})(s) + \mathbf{C}\mathbf{R}(s)\mathbf{x}(0)$$

for all $s \in \mathbb{C}$, where

$$\mathcal{H}(s) \;:=\; \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}$$

is called the *transfer function* of Σ, and

$$\mathbf{R}(s) \;:=\; (s\mathbf{I} - \mathbf{A})^{-1}.$$

Assume now that $\mathbf{x}_0 := \mathbf{0}$ and that the systems Σ and $\widehat{\Sigma}$ have identical input $\mathbf{u} \in \mathbb{U}$. Then in requiring that $\mathcal{H}|_{i\Omega} = \widehat{\mathcal{H}}|_{i\Omega}$ on some set of frequencies $\omega \in \Omega \subset \mathbb{R}$, it follows that $(\mathcal{L}\mathbf{y})|_{i\Omega} = (\mathcal{L}\widehat{\mathbf{y}})|_{i\Omega}$ because

$$(\mathcal{L}\mathbf{y})(i\omega) - (\mathcal{L}\widehat{\mathbf{y}})(i\omega) = \Big(\mathcal{H}(i\omega) - \widehat{\mathcal{H}}(i\omega)\Big)(\mathcal{L}\mathbf{u})(i\omega). \tag{1.3}$$

Moment matching techniques approximate the output of the full-order model Σ with

that of the reduced-order model $\widehat{\Sigma}$ by requiring that $\widehat{\mathcal{H}} \approx \mathcal{H}$, which is attained by matching terms in the series expansions for $\mathcal{H}$ and $\widehat{\mathcal{H}}$ at frequencies near some critical value $i\sigma \in i\mathbb{R}$. In attaining that objective, one finds that the output error

$$\begin{aligned} \mathbf{y}(t) - \widehat{\mathbf{y}}(t) &= \tfrac{1}{2\pi} \int_{\mathbb{R}} \Big((\mathcal{L}\mathbf{y})(i\omega) - (\mathcal{L}\widehat{\mathbf{y}})(i\omega) \Big) e^{i\omega t} d\omega \\ &= \tfrac{1}{2\pi} \int_{\mathbb{R}} \Big(\mathcal{H}(i\omega) - \widehat{\mathcal{H}}(i\omega) \Big) (\mathcal{L}\mathbf{u})(i\omega) e^{i\omega t} d\omega \end{aligned} \tag{1.4}$$

is small.

Expansion of $\mathcal{H}$ About 0

Assume that $s \in \mathbb{C}$ is adequately close to 0, so that $\mathcal{H}$ can be expanded in the convergent series

$$\mathcal{H}(s) = \mathbf{H}_0 + s\mathbf{H}_1 + s^2\mathbf{H}_2 + \cdots ,$$

where $\{\mathbf{H}_j\}_{j=0}^{\infty} \subseteq \mathbb{C}^{p \times m}$ are the *moments of* $\mathcal{H}$ *expanded at 0*. One can expect to attain $\widehat{\mathcal{H}}(s) \approx \mathcal{H}(s)$ by finding a reduced system $\widehat{\Sigma}$ that satisfies

$$\widehat{\mathcal{H}}(s) = \mathbf{H}_0 + s\mathbf{H}_1 + s^2\mathbf{H}_2 + \cdots + s^q\mathbf{H}_q + s^{q+1}\widehat{\mathbf{H}}_{q+1} + s^{q+2}\widehat{\mathbf{H}}_{q+2} + \cdots .$$

This problem of *matching the first* $q \in \mathbb{N}$ *moments* of $\mathcal{H}$ with the first q moments of $\widehat{\mathcal{H}}$ amounts to finding $\widehat{\Sigma}$ such that

$$\widehat{\mathbf{H}}_j = \mathbf{H}_j$$

for $j \in \{0, \ldots, q-1\}$ [1, p. 88, 97, 343-346].

Without loss of generality, assume $\mathbf{D} = \mathbf{0}$ (the reduced system would otherwise contain $\widehat{\mathbf{D}} := \mathbf{D}$) and consider the series expansion for $\mathcal{H}$ around $0 \in \mathbb{C}$. That is, for $s \in \mathrm{B}_{\frac{1}{\rho(\mathbf{A}^{-1})}}(0)$ — the ball of radius $1/\rho(\mathbf{A}^{-1})$ around 0 — it holds that

$$\begin{aligned}
\mathcal{H}(s) &= \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} \\
&= -\mathbf{C}(\mathbf{I} - s\mathbf{A}^{-1})^{-1}\mathbf{A}^{-1}\mathbf{B} \\
&= -\mathbf{C}\sum_{j=0}^{\infty}(s\mathbf{A}^{-1})^{j}\mathbf{A}^{-1}\mathbf{B} \\
&= \sum_{j=0}^{\infty} s^{j}\mathbf{H}_{j},
\end{aligned}$$

where

$$\mathbf{H}_j := -\mathbf{C}\mathbf{A}^{-(j+1)}\mathbf{B} \tag{1.5}$$

are the moments of $\mathcal{H}$ for frequencies $s \in \mathbb{C}$ near 0; see [7, Chapter 2].

Then matching the first q moments of Σ near 0 with the first q moments of the reduced $\widehat{\Sigma}$ near 0 amounts to finding $\widehat{\mathbf{A}} \in \mathbb{C}^{\widehat{n}\times\widehat{n}}$, $\widehat{\mathbf{B}} \in \mathbb{C}^{\widehat{n}\times m}$ and $\widehat{\mathbf{C}} \in \mathbb{C}^{p\times\widehat{n}}$ that satisfy $\widehat{\mathbf{H}}_j = \mathbf{H}_j$ for all $j \in \{0, \ldots, q-1\}$, i.e.,

$$\widehat{\mathbf{C}}\widehat{\mathbf{A}}^{-(j+1)}\widehat{\mathbf{B}} = \mathbf{C}\mathbf{A}^{-(j+1)}\mathbf{B}.$$

In finding such $\widehat{\Sigma}$, one expects that $(\mathcal{L}\mathbf{y})|_{\left\{|s|<\frac{1}{\rho(\mathbf{A}^{-1})}\right\}} \approx (\mathcal{L}\widehat{\mathbf{y}})|_{\left\{|s|<\frac{1}{\rho(\mathbf{A}^{-1})}\right\}}$ on this set of low frequencies, i.e., frequencies close to 0. Such an approach is logical if $\mathbf{y}$ is

dominated by low frequencies.

1.2.3 Moments of $\mathcal{H}$ About Arbitrary $\sigma \in \mathbb{C}$ and About $+\infty$

Analogously to matching moments on a set of low frequencies $\left\{|s| < \frac{1}{\rho(\mathbf{A}^{-1})}\right\} \subset \mathbb{C}$ by expanding $\mathcal{H}$ at frequencies near 0, one can also match moments on some set Ω near some arbitrary frequency $\sigma \in \mathbb{C}$ or on some set Ω "in a neighborhood of infinity," i.e., frequencies of arbitrarily large magnitudes.

Series Expansion for $\mathcal{H}$ About Arbitrary $\sigma \in \mathbb{C}$

Consider first an expansion for $\mathcal{H}$ near some fixed frequency $\sigma \in \mathbb{C}$. Define $\mathbf{A}_\sigma := (\mathbf{A} - \sigma\mathbf{I})$. Require $|s| < 1/\rho(\mathbf{A}_\sigma^{-1})$ and observe

$$\begin{aligned}
\mathcal{H}(\sigma + s) &= \mathbf{C}\left[(\sigma + s)\mathbf{I} - \mathbf{A}\right]^{-1}\mathbf{B} \\
&= \mathbf{C}\left[s\mathbf{I} - \mathbf{A}_\sigma\right]^{-1}\mathbf{B} \\
&= -\mathbf{C}\left[\mathbf{I} - s\mathbf{A}_\sigma^{-1}\right]^{-1}\mathbf{A}_\sigma^{-1}\mathbf{B} \\
&= -\mathbf{C}\sum_{j=0}^{\infty}(s\mathbf{A}_\sigma^{-1})^j\mathbf{A}_\sigma^{-1}\mathbf{B} \\
&= \sum_{j=0}^{\infty}\mathbf{H}_j s^j,
\end{aligned}$$

where $\mathbf{H}_j := -\mathbf{C}\mathbf{A}_\sigma^{-(j+1)}\mathbf{B}$ are moments of $\mathcal{H}$ at frequencies $\sigma + s$ near $\sigma \in \mathbb{C}$. In finding $\widehat{\Sigma}$ such that the first q moments match, i.e., for all $j \in \{0, \ldots, q-1\}$,

$$\widehat{\mathbf{C}}\widehat{\mathbf{A}}_\sigma^{-(j+1)}\widehat{\mathbf{B}} = \mathbf{C}\mathbf{A}_\sigma^{-(j+1)}\mathbf{B},$$

one expects that $(\mathcal{L}\mathbf{y})|_\Omega \approx (\mathcal{L}\widehat{\mathbf{y}})|_\Omega$ on the set of frequencies $\Omega := \mathrm{B}_{\frac{1}{\rho(\mathbf{A}_\sigma^{-1})}}(\sigma) \subset \mathbb{C}$ close to σ [7, Chapter 2]. Such an approach is logical if $\mathbf{y}$ is dominated by frequencies near σ.

Series Expansion for $\mathcal{H}$ About $+\infty$

To expand $\mathcal{H}$ in a neighborhood of $+\infty$, i.e., at frequencies s of arbitrarily large magnitude, assume that $|s| > \rho(\mathbf{A})$ and observe that

$$\begin{aligned}
\mathcal{H}(s) &= \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} \\
&= (1/s)\mathbf{C}\Big(\mathbf{I} - (1/s)\mathbf{A}\Big)^{-1}\mathbf{B} \\
&= (1/s)\mathbf{C}\sum_{j=0}^{\infty}\Big((1/s)\mathbf{A}\Big)^{j}\mathbf{B} \\
&= \sum_{j=0}^{\infty}(1/s^{j+1})\mathbf{H}_j,
\end{aligned}$$

where $\mathbf{H}_j := \mathbf{C}\mathbf{A}^j\mathbf{B}$ are the moments of $\mathcal{H}$ in a neighborhood of $+\infty$, also called the *Markov parameters* of Σ. Analogously to the expansions around 0 and $\sigma \in \mathbb{C}$, matching the first q moments $\mathbf{H}_j = \widehat{\mathbf{H}}_j$ near $+\infty$ amounts to finding $\widehat{\Sigma}$ such that for

all $j \in \{0, \ldots, q-1\}$,

$$\widehat{\mathbf{C}}\widehat{\mathbf{A}}^j\widehat{\mathbf{B}} = \mathbf{C}\mathbf{A}^j\mathbf{B}$$

[7, Chapter 2]. In matching moments at $+\infty$, one expects that $(\mathcal{L}\mathbf{y})|_{\{|s|>\rho(\mathbf{A})\}} \approx (\mathcal{L}\widehat{\mathbf{y}})|_{\{|s|>\rho(\mathbf{A})\}}$. Such an approach is logical if $\mathbf{y}$ is dominated by high frequencies.

1.2.4 Moment Matching Using Arnoldi

Consider now the single input single output (*SISO*) system

$$\Sigma := \left(\begin{array}{cc} \mathbf{A} & \mathbf{b} \\ \mathbf{c} & \mathbf{0} \end{array} \right), \tag{1.6}$$

where

$$\begin{array}{rcl} \mathbf{A} & \in & \mathbb{C}^{n\times n}, \\ \mathbf{b} & \in & \mathbb{C}^{n}, \\ \mathbf{c} & \in & \mathbb{C}^{1\times n}. \end{array} \tag{1.7}$$

Define the kth Krylov subspace by

$$\mathcal{K}_k(\mathbf{A}, \mathbf{b}) \;:=\; \text{span}\left\{\mathbf{b}, \mathbf{A}\mathbf{b}, \ldots, \mathbf{A}^{k-1}\mathbf{b}\right\}.$$

Arnoldi's method applied to $(\mathbf{A}, \mathbf{b})$ iteratively generates $\mathbf{Q}_k \in \mathbb{C}^{n\times k}$, $\mathbf{H}_k \in \mathbb{C}^{k\times k}$, $\mathbf{q}_{k+1} \in \mathbb{C}^n$ and $h_{k+1,k} \in \mathbb{R}$ such that

$$\mathbf{A}\mathbf{Q}_k = \mathbf{Q}_k\mathbf{H}_k + h_{k+1,k}\mathbf{q}_{k+1}\mathbf{e}_k^T,$$

where the columns of $\mathbf{Q}_k$ form an orthonormal basis for $\mathcal{K}_k(\mathbf{A}, \mathbf{b})$, $\mathbf{H}_k$ is upper Hessenberg, and $\mathbf{Q}_k^* \mathbf{q}_{k+1} = 0$. When $\mathbf{A}$ is Hermitian, the Arnoldi process simplifies to the *Lanczos iteration (see, e.g.* [21, Lecture 36]).

*Theorem 1.1 (**Arnoldi Matches Moments**)*

(i) *Suppose that* $\mathbf{A}\mathbf{Q}_k = \mathbf{Q}_k \mathbf{H}_k + h_{k+1,k} \mathbf{q}_{k+1} \mathbf{e}_k^T$ *is an Arnoldi factorization for the space* $\mathcal{K}_k(\mathbf{A}, \mathbf{b})$. *Then the reduced system*

$$\widehat{\Sigma} := \begin{pmatrix} \widehat{\mathbf{A}} & \widehat{\mathbf{b}} \\ \widehat{\mathbf{c}} & \mathbf{0} \end{pmatrix} := \begin{pmatrix} \mathbf{Q}_k^* \mathbf{A} \mathbf{Q}_k & \mathbf{Q}_k^* \mathbf{b} \\ \mathbf{c} \mathbf{Q}_k & \mathbf{0} \end{pmatrix} \tag{1.8}$$

matches the first k *Markov parameters of* Σ. *That is, for all* $j \in \{0, \ldots, k-1\}$,

$$\widehat{\mathbf{c}} \widehat{\mathbf{A}}^j \widehat{\mathbf{b}} \;=\; \mathbf{c} \mathbf{A}^j \mathbf{b}.$$

(ii) *Suppose that* $\mathbf{A}^{-1}\mathbf{Q}_k = \mathbf{Q}_k \mathbf{H}_k + h_{k+1,k} \mathbf{q}_{k+1} \mathbf{e}_k^T$ *is an Arnoldi factorization for the space* $\mathcal{K}_k(\mathbf{A}^{-1}, \mathbf{b})$. *Then the reduced system*

$$\widehat{\Sigma} := \begin{pmatrix} \widehat{\mathbf{A}} & \widehat{\mathbf{b}} \\ \widehat{\mathbf{c}} & \mathbf{0} \end{pmatrix} := \begin{pmatrix} \mathbf{Q}_k^* \mathbf{A} \mathbf{Q}_k & \mathbf{Q}_k^* \mathbf{b} \\ \mathbf{c} \mathbf{Q}_k & \mathbf{0} \end{pmatrix} \tag{1.9}$$

matches the first k *moments of* Σ *and* $\widehat{\Sigma}$ *at* 0. *That is, for all* $j \in \{0, \ldots, k-1\}$,

$$\widehat{\mathbf{c}} \widehat{\mathbf{A}}^{-j} \widehat{\mathbf{b}} \;=\; \mathbf{c} \mathbf{A}^{-j} \mathbf{b}.$$

See for example, [7, Chapter 2], for a more detailed explanation.

1.3 Dimension Reduction by Modal Filtering

1.3.1 Derivation of Davison's Method for Diagonalizable $\mathbf{A} \in \mathbb{C}^{n\times n}$

This initial derivation of Davison's Method is based largely on the exposition of Bonvin and Mellichamp [2].

Consider the system

$$\Sigma_1 := \left(\begin{array}{lcl} \dot{\mathbf{x}} & = & \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \quad \text{for } t \geq 0 \\ \mathbf{x}(0) & = & \mathbf{x}_0 \end{array} \right) \tag{1.10}$$

with $\mathbf{u} \in \mathbb{U}$ and $\mathbf{x} \in \mathbb{X}$ ($\mathbf{A} \in \mathbb{C}^{n\times n}$, $\mathbf{B} \in \mathbb{C}^{n\times m}$). Modal filtering involves identifying and discarding those least important eigenmodes of $\mathbf{A} \in \mathbb{R}^{n\times n}$ from the system Σ_1. Assume that $\mathbf{A}$ is diagonalizable, and — for the purpose of illustration — that n is adequately small to justify the computation of the diagonalization $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$. Assume further that the modes of $\mathbf{A}$ are already ordered from most to least important (according to a ranking of choice) within this factorization. If this is not the case, obtain a row permutation matrix $\mathbf{R}$ to re-order the modes as desired and redefine

$$\begin{array}{rcl} \mathbf{\Lambda} & := & \mathbf{R}^T\mathbf{\Lambda}\mathbf{R}, \\ \mathbf{V} & := & \mathbf{V}\mathbf{R}^T, \end{array} \tag{1.11}$$

so that $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$ is properly ordered.

Set $\mathbf{z} := \mathbf{V}^{-1}\mathbf{x} \in \mathbb{X}$ and observe that Σ_1 is equivalent to Σ_2 given by

$$\Sigma_2 \;:=\; \left(\begin{array}{lcl} \dot{\mathbf{z}} & = & \mathbf{\Lambda z} + \mathbf{V}^{-1}\mathbf{B u} \quad \text{for } t \geq 0 \\ \mathbf{z}(0) & = & \mathbf{V}^{-1}\mathbf{x}_0 \end{array} \right).$$

Notice that all modes are decoupled in Σ_2. One can identify a *truncation node* $l \in \{1, \ldots, n\}$ that separates the important and unimportant equations in Σ_2 into two systems —

$$\begin{array}{rcll} \dot{\mathbf{z}_1} & = & \mathbf{\Lambda}_1\mathbf{z}_1 + \mathbf{W}_1^*\mathbf{B u} & \text{for } t \geq 0, \\ \dot{\mathbf{z}_2} & = & \mathbf{\Lambda}_1\mathbf{z}_2 + \mathbf{W}_2^*\mathbf{B u} & \text{for } t \geq 0, \end{array} \tag{1.12}$$

where I have partitioned

$$\mathbf{z} = \left(\begin{array}{c} \mathbf{z}_1 \\ \mathbf{z}_2 \end{array} \right)$$

with $\mathbf{z}_1 \in \mathbb{C}^l$ and $\mathbf{z}_2 \in \mathbb{C}^{n-l}$,

$$\mathbf{V} \;=\; \left(\begin{array}{cc} \mathbf{V}_1 & \mathbf{V}_2 \end{array} \right),$$

and

$$\mathbf{V}^{-1} \;=\; \left(\begin{array}{c} \mathbf{W}_1^* \\ \mathbf{W}_2^* \end{array} \right).$$

Now one can weigh the relevance of the modes in $\mathbf{z}$ through various means. Davison's method does so by assuming that the contribution due to $\mathbf{z}_2$ is $\mathbf{0}$ at all times.

Note that symbolically this is not correct because $\mathbf{W}_2^*\mathbf{B}\mathbf{u} \neq 0$ for all times $t \geq 0$ in general. Nonetheless, it is logical to neglect the contribution made by $\mathbf{z}_2$ by considering only the first set of equations in (1.12). That is, approximate $\mathbf{z}$ by

$$\mathbf{z} \approx \widetilde{\mathbf{z}} = \begin{pmatrix} \widetilde{\mathbf{z}}_1 \\ \widetilde{\mathbf{z}}_2 \end{pmatrix} := \begin{pmatrix} \mathbf{z}_1 \\ \mathbf{0} \end{pmatrix}$$

for all $t \in \mathbb{R}$, where $\widetilde{\mathbf{z}}_1 = \mathbf{z}_1$. (The assumption $\mathbf{z}_2 \approx \mathbf{0}$ does not affect the value of $\mathbf{z}_1$, since the modes contained in the blocks $\mathbf{z}_1$ and $\mathbf{z}_2$ are decoupled.) This leads to a further refined system Σ_3 given by

$$\Sigma_3 := \begin{pmatrix} \dot{\mathbf{z}}_1 & = & \mathbf{\Lambda}_1\mathbf{z}_1 + \mathbf{W}_1^*\mathbf{B}\mathbf{u} \quad \text{for } t \geq 0 \\ \mathbf{z}_1(0) & = & \mathbf{W}_1^*\mathbf{x}_0 \end{pmatrix}, \tag{1.13}$$

where I have partitioned

$$\mathbf{\Lambda} = \begin{pmatrix} \mathbf{\Lambda}_1 & \\ & \mathbf{\Lambda}_2 \end{pmatrix}.$$

Observe then that solving Σ_3 yields $\widetilde{\mathbf{x}} : \mathbb{R} \to \mathbb{C}^n$ that approximates $\mathbf{x} : \mathbb{R} \to \mathbb{C}^n$, because at times $t \in \mathbb{R}$,

$$\begin{aligned} \mathbf{x} &= \mathbf{V}\mathbf{z} \\ &\quad (\Sigma_1 \to \Sigma_2) \\ &\approx \mathbf{V}\widetilde{\mathbf{z}} \\ &\quad (\Sigma_2 \to \Sigma_3) \\ &= \begin{pmatrix} \mathbf{V}_1 & \mathbf{V}_2 \end{pmatrix} \begin{pmatrix} \mathbf{z}_1 \\ \mathbf{0} \end{pmatrix} \\ &= \mathbf{V}_1\mathbf{z}_1 \\ &=: \widetilde{\mathbf{x}}. \end{aligned}$$

As Σ_3 contains only relevant modes of $\mathbf{A}$, which are generally low frequency modes, it is more quickly integrated stably using forward Euler than the original system Σ_1. It is also of dimension $l \ll n$, so that computations with Σ_3 are faster than those using Σ_1. Moreover, under the assumption that the high frequency modes contribute negligibly to the exact solution $\mathbf{x}$ to Σ_1 (as is often the case, as Chapter 2 will show), approximating $\mathbf{x}$ using $\widetilde{\mathbf{x}}$ obtained through Σ_3 yields little loss of accuracy.

Notice that given some output $\mathbf{y} : \mathbb{R} \to \mathbb{C}^p$ in Σ_1 defined by

$$\Sigma_1 := \left(\begin{array}{lcll} \dot{\mathbf{x}} & = & \mathbf{Ax} + \mathbf{Bu} & \text{for } t \geq 0 \\ \mathbf{y} & = & \mathbf{Cx} + \mathbf{Du} & \text{for } t \geq 0 \\ \mathbf{x}(0) & = & \mathbf{x}_0 & \end{array} \right),$$

one applies Davison's method in the exact same manner and obtains the approximation $\widetilde{\mathbf{y}} \approx \mathbf{y}$ at all $t \in \mathbb{R}$ through

$$\widetilde{\mathbf{y}} := \mathbf{C}\widetilde{\mathbf{x}} + \mathbf{Du}. \tag{1.14}$$

1.3.2 Viewing Modal Filtering as a Projection

I now deviate from Bonvin and Mellichamp in my discussion of modal reduction by showing that Davison's method is simply a projection of the exact full-sized solution $\mathbf{x}$ onto a subspace $\mathcal{U}$ of $\mathbb{C}^n$.

Recall from Section 1.3.1 that reducing Σ_1 to Σ_3 using Davison's technique amounts to finding $\widetilde{\mathbf{x}} : \mathbb{R} \to \mathbb{C}^n$ that approximates $\mathbf{x} : \mathbb{R} \to \mathbb{C}^n$ and satisfies $\widetilde{\mathbf{x}} = \mathbf{V}_1\mathbf{z}_1$ and

$\mathbf{z} = \mathbf{V}^{-1}\mathbf{x}$ for all times $t \in \mathbb{R}$. Hence

$$\widetilde{\mathbf{x}} \;=\; \mathbf{V}_1\mathbf{W}_1^*\mathbf{x}. \tag{1.15}$$

Now denote the rows of $\mathbf{V}^{-1}$ by $\mathbf{w}_i^* \in \mathbb{C}^{1\times n}$, i.e.,

$$\mathbf{V}^{-1} = \begin{pmatrix} \mathbf{w}_1^* \\ \vdots \\ \mathbf{w}_n^* \end{pmatrix}.$$

Thus

$$\begin{aligned} \mathbf{V}_1\mathbf{W}_1^* &= \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_l \end{pmatrix} \begin{pmatrix} \mathbf{w}_1^* \\ \vdots \\ \mathbf{w}_l^* \end{pmatrix} \\ &= \sum_{i=1}^{l} \mathbf{v}_i\mathbf{w}_i^* \\ &= \sum_{i=1}^{l} \mathbf{P}_i, \end{aligned}$$

where $\mathbf{P}_i := \mathbf{v}_i\mathbf{w}_i^*$ is the projector onto the span of $\mathbf{v}_i$. Yet

$$\sum_{i=1}^{l} \mathbf{P}_i \;=\; \mathbf{P}_{\mathcal{U}},$$

the projection $\mathbf{P}_{\mathcal{U}} : \mathbb{C}^n \to \mathbb{C}^n$ onto the subspace $\mathcal{U} := \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_l\}$.

Hence the approximation $\widetilde{\mathbf{x}} : \mathbb{R} \to \mathbb{C}^n$ to $\mathbf{x} : \mathbb{R} \to \mathbb{C}^n$ obtained through Davison's method is simply a projection of $\mathbf{x} : \mathbb{R} \to \mathbb{C}^n$ onto the space spanned by the most important eigenvectors $\{\mathbf{v}_i\}_{i=1}^{l}$ of $\mathbf{A}$, that is

$$\widetilde{\mathbf{x}} \;=\; \mathbf{V}_1\mathbf{W}_1^*\mathbf{x} \;=\; \mathbf{P}_{\mathcal{U}}\mathbf{x}. \tag{1.16}$$

Of course, one would never implement Davison's method using equation (1.16) because to do so would defeat the purpose of the modal reduction (for using equation (1.16) requires integration of the full-sized system Σ_1).

1.3.3 Additional Comments on Modal Filtering

Note that Davison's method generally becomes increasingly impractical for systems of the form (1.2) as n increases because $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$ is increasingly costly to compute and store. In particular, when $\mathbf{A}$ is sparse, $\mathbf{V}$ is not generally sparse, so one often loses the advantage of sparse storage of $\mathbf{A}$ when storing the corresponding $\mathbf{V}$. Note that one can compute a subset of $\sigma(\mathbf{A})$ (with corresponding right and left eigenvectors) in these circumstances through use of ARPACK/`Matlab`'s `eigs`, routines that capitalize on sparse structure of $\mathbf{A}$.

Notice also that Davison's method makes an error in its approximation immediately at time $t = 0$ through performing

$$\mathbf{x}(0) \approx \widetilde{\mathbf{x}}(0) := \mathbf{W}_1^*\mathbf{x}_0.$$

Error in the approximation at time 0 is undesirable. Nonetheless, performing this approximation at time 0 ensures that the approximate solution $\widetilde{\mathbf{x}} : \mathbb{R} \to \mathbb{R}^n$ is not

discontinuous when restricted to the non-negative portion of the time domain, $[0, \infty)$, due to an instantaneous disappearance of unimportant modes immediately following time 0 when the projector $\mathbf{P}_{\mathcal{U}}$ becomes active. It also insures that the initial condition $\widetilde{\mathbf{x}}(0)$ is not influenced by irrelevant eigenmodes.

Chapter 2

The Heat Equation

To understand how the various eigenmodes of the full-sized system (1.2) influence the behavior of a reduced-order model, particularly one of the form (1.8), I turn to the heat equation. The spatial discretization of the heat equation yields linear, time-invariant systems in which the high frequency modes — those that restrict the stable time-step for the forward Euler scheme — are nearly insignificant. Ideally, in such a context, all influence by the high frequency modes would be ignored, yet the system dimension is not generally small enough to justify modal filtration, as such requires diagonalization of the matrix $\mathbf{A}$. Yet after taking the preliminary step of dimension reduction via moment matching, modal filtration becomes computationally palatable. In addition to providing an exposition on the decay of coefficients in discrete Fourier sine series (I occasionally abbreviate by *FSS*) expansions, Chapter 2 unveils a criterion for selecting a cut-off point for the modal truncation step in a dual-stage procedure whose first reduction stage is performed by moment matching. This dual-stage technique works remarkably well for semi-discretized heat equations.

2.1 Solution to the Continuous Heat Equation

In sections 2.1.1 and 2.1.2, I rely heavily on Gockenbach's *Partial Differential Equations: Analytical and Numerical Methods* [10].

2.1.1 Eigenvalues and Eigenfunctions of $\mathrm{L} := \triangle$

Fix an arbitrary length $\beta \in \mathbb{R}$, and define the domain $\Omega := [0, \beta]$. Consider the function space

$$\mathcal{Q} := \left\{ g \in C^2(\Omega) : g|_{\partial\Omega} = 0 \right\} \tag{2.1}$$

with the associated inner product $\langle \cdot, \cdot \rangle_\Omega : \mathrm{L}^2(\Omega) \times \mathrm{L}^2(\Omega) \to \mathbb{R}$ given by

$$\langle g, h \rangle_\Omega := \int_\Omega g(x)\overline{h(x)}dx.$$

Define the differential operator $\mathrm{L} : \mathcal{Q} \to \mathrm{L}^2(\Omega)$ by

$$\mathrm{L}g := \triangle g \tag{2.2}$$

and observe that L is *self-adjoint* in $\langle \cdot, \cdot \rangle_\Omega$, i.e., $\langle \mathrm{L}f, g \rangle_\Omega = \langle f, \mathrm{L}g \rangle_\Omega$ for all $f,\ g \in \mathcal{Q}$. The eigenfunctions of L,

$$\phi_j(x) := \sqrt{\tfrac{2}{\beta}} \sin\left(\tfrac{\pi j x}{\beta}\right), \quad j \in \mathbb{N}, \tag{2.3}$$

form an orthonormal basis for $\mathcal{Q}$. Hence any $g \in \mathcal{Q}$ can be represented using its *Fourier sine series*,

$$g(x) = \sum_{j \in \mathbb{N}} s_j \phi_j(x), \tag{2.4}$$

where $s_j := \langle g, \phi_j \rangle_\Omega$ is the jth *Fourier sine series coefficient* of g. The eigenvalue corresponding to ϕ_j is

$$\lambda_j := -\left(\frac{\pi j}{\beta}\right)^2. \tag{2.5}$$

See, e.g., [10, pages 136-144].

2.1.2 Fourier Series Expansion

Fix some final time $T > 0$. Given $f : \Omega \times (-\infty, T] \to \mathbb{R}$, $f|_{\Omega \times (-\infty, 0)} = 0$, and $w_0 : \Omega \to \mathbb{R}$, consider the Dirichlet initial boundary value problem (IBVP) given by

$$\begin{aligned} w_t &= \mathrm{L}w + f \quad on \quad \Omega \times [0, T], \\ w|_{\Omega \times \{0\}} &= w_0, \\ w|_{\partial\Omega \times [0,T]} &= 0, \\ w|_{\Omega \times (-\infty, 0)} &= 0. \end{aligned} \tag{2.6}$$

To express the solution $w : \Omega \times (-\infty, T] \to \mathbb{R}$ as a Fourier series, expand $f : \Omega \times (-\infty, T] \to \mathbb{R}$ in its time-dependent series

$$f(x,t) \;=\; \sum_{j \in \mathbb{N}} \widetilde{s}_j(t) \phi_j(x),$$

where $\widetilde{s}_j(t) := \langle f(\cdot, t), \phi_j \rangle_\Omega$ for $t \geq 0$, and $\widetilde{s}_j(t) := 0$ for $t < 0$, and solve for w of the

form

$$w(x,t) = \sum_{j\in\mathbb{N}} s_j(t)\phi_j(x), \tag{2.7}$$

where $s_j(t) := \langle w(\cdot,t), \phi_j\rangle_\Omega$ for $t \geq 0$, and $s_j(t) := 0$ for $t < 0$. Immediately then, $s_j : (-\infty, T] \to \mathbb{R}$ is derived at non-negative times by observing that $w_t = \mathrm{L}w + f$ implies

$$s_j'(t) = \langle w_t, \phi_j\rangle_\Omega = \langle \mathrm{L}w + f, \phi_j\rangle_\Omega = \lambda_j s_j(t) + \widetilde{s}_j(t),$$

which has the well-known solution

$$s_j(t) = e^{\lambda_j t} s_j(0) + \int_0^t e^{(t-r)\lambda_j}\widetilde{s}_j(r)dr \tag{2.8}$$

for $t \geq 0$. (See, e.g., [9, eq. 4.1], [10, p. 195-197].) In particular, when $w_0 = 0$,

$$s_j(t) \;=\; \int_0^t e^{(t-r)\lambda_j}\widetilde{s}_j(r)dr$$

for $t \geq 0$.

2.2 Fourier Series Coefficient Decay

The expansion of a function in the basis formed by the Laplacian's eigenfunctions can be regarded as the Fourier series of the odd extension of a function to the domain $\Omega_2 := [-\beta, \beta]$. In order to understand the decay of coefficients in Fourier sine

series expansions, one must first explore the decay of coefficients in full Fourier series expansions. To this end, Section 2.2.1 will focus on the decay of complex Fourier series coefficients before focusing on the decay of Fourier sine series coefficients in Section 2.2.2.

2.2.1 Complex Fourier Series for $f : \Omega_2 \to \mathbb{R}$

Take $\langle \cdot, \cdot \rangle_{\Omega_2} : \mathrm{L}^2(\Omega_2) \times \mathrm{L}^2(\Omega_2) \to \mathbb{R}$ to be the $\mathrm{L}^2(\Omega_2)$ inner product.

Any piecewise smooth $f : \Omega_2 \to \mathbb{R}$ whose periodic extension is also piecewise smooth can be represented by its *complex Fourier series*,

$$s(x) \;:=\; \textstyle\sum_{j\in\mathbb{Z}} c_j \Phi_j(x), \tag{2.9}$$

where $\Phi_j(x) := e^{-i\pi jx/\beta}$, in the sense that

$$\|f - s\|_{\mathrm{L}^2(\Omega_2)} = 0, \tag{2.10}$$

where the jth *complex Fourier series coefficient* is denoted by

$$c_j := \frac{1}{2\beta} \left\langle f, e^{i\pi jx/\beta} \right\rangle_{\Omega_2} =: \frac{1}{2\beta} \left\langle f, \overline{\Phi_j} \right\rangle_{\Omega_2}, \tag{2.11}$$

[10, p. 537], [4, p. 37].

The *real Fourier series* of $f : \Omega_2 \to \mathbb{R}$ is given by

$$f(x) = a_0 + \sum_{j \in \mathbb{N}} \left(a_j \cos\left(\frac{j\pi x}{\beta}\right) + b_j \sin\left(\frac{j\pi x}{\beta}\right) \right), \tag{2.12}$$

where, for $j \geq 1$, [10, p. 537]

$$\begin{aligned} a_0 &:= \frac{1}{2\beta} \langle f, 1 \rangle_{\Omega_2}, \\ a_j &:= \frac{1}{\beta} \left\langle f, \cos\left(\frac{j\pi x}{\beta}\right) \right\rangle_{\Omega_2}, \\ b_j &:= \frac{1}{\beta} \left\langle f, \sin\left(\frac{j\pi x}{\beta}\right) \right\rangle_{\Omega_2}. \end{aligned}$$

*Theorem 2.1 (**Equivalent Formula for Complex Fourier Series Coefficients**) Suppose that for some $p \geq 2$, the periodic extension of $f : \Omega_2 \to \mathbb{R}$ given by $f_{per} : \mathbb{R} \to \mathbb{R}$ is $C^{p-2}(\mathbb{R})$ and that $f_{per}^{(p-1)} : \mathbb{R} \to \mathbb{R}$ (the periodic extension of $f^{(p-1)} : \Omega_2 \to \mathbb{R}$ to all of $\mathbb{R}$) is piecewise smooth and bounded on $\mathbb{R}$. Then the jth complex Fourier series coefficient given by $c_j := \frac{1}{2\beta} \left\langle f, \overline{\Phi_j} \right\rangle_{\Omega_2}$ satisfies*

$$c_j = \left(\frac{1}{2\beta}\right) \left(\frac{l}{i\pi j}\right)^{p-1} \left\langle f^{(p-1)}, \overline{\Phi_j} \right\rangle_{\Omega_2}.$$

(See, e.g., [10, Theorem 12.20] and [4, Theorem 6.1].)

For a proof, see Appendix A.1.

*Lemma 2.1 (**Corollary: Decay of Complex Fourier Series Coefficients**)*
Under the assumptions of the previous theorem,

$$|c_j| \;\leq\; \left(\frac{\beta}{\pi|j|}\right)^{p-1} \|f^{(p-1)}\|_{\mathrm{L}^\infty(\Omega_2)}.$$

Hence, as $j \to \infty$

$$|c_j| \;=\; \mathcal{O}\left(\,|j|^{-(p-1)}\,\right).$$

(See, e.g., [10, Theorem 12.20] and [4, Theorem 6.1].)

Proof

Observe that $\left|\int_{\Omega_2} f^{(p-1)}(x)\Phi_j(x)dx\right| \leq 2\beta\|f^{(p-1)}\|_{\mathrm{L}^\infty(\Omega_2)}$. Hence

$$\begin{aligned} |c_j| &= \left|\left(\frac{1}{2\beta}\right)\left(\frac{\beta}{i\pi j}\right)^{p-1}\left\langle f^{(p-1)}, \overline{\Phi_j}\right\rangle_{\Omega_2}\right| \\ &\leq \left(\frac{\beta}{\pi|j|}\right)^{p-1}\|f^{(p-1)}\|_{\mathrm{L}^\infty(\Omega_2)}. \end{aligned}$$

♦

When $f : \Omega_2 \to \mathbb{R}$ is odd, Theorem 2.1 can be improved to provide better estimates on the decay of complex Fourier series coefficients when $f^{(p-1)}$ or $f^{(p)}$ is unbounded at the singleton points in the set $\{0,\ \pm a,\ \pm\beta\}$ for some $a \in (0, \beta)$.

*Theorem 2.2 (**Complex Coefficient Decay with Singularities at** $0,\ \pm a,\ \pm\beta$)*
Consider an odd function, $f : \Omega_2 \to \mathbb{R}$*, with its corresponding periodic extension given by* $f_{per} : \mathbb{R} \to \mathbb{R}$*. Recall that the* j*th complex Fourier series coefficient is given*

by $c_j := \frac{1}{2\beta} \left\langle f, \overline{\Phi_j} \right\rangle_{\Omega_2}$.

(i) *Suppose that for some* $p \geq 2$, $f_{per} \in C^{p-2}(\mathbb{R})$, *and* $f_{per}^{(p-1)} : \mathbb{R} \to \mathbb{R}$ *is piecewise continuous on* $\mathbb{R}$ *and bounded on* $\mathbb{R} \setminus \{0,\ z\beta,\ za\}_{z\in\mathbb{Z}}$.

Then c_j *satisfies*

$$|c_j| \ \leq \ \frac{\beta^{p-2}}{(\pi|j|)^{p-1}} \left(\begin{array}{l} \beta \| f^{(p-1)} \|_{\mathrm{L}^\infty(\Omega_2 \setminus \{0,\ \pm a,\ \pm\beta\})} \\ \\ +|f^{(p-2)}(0)| + 2|f^{(p-2)}(a)| + 2|f^{(p-2)}(\beta)| \end{array} \right).$$

Hence as $j \to \infty$,

$$|c_j| \ = \ \mathcal{O}\big(\ |j|^{-(p-1)} \ \big).$$

(ii) *Alternatively, suppose that for some* $p \geq 2$, $f_{per} \in C^{p-2}(\mathbb{R})$, $f_{per}^{(p-1)} : \mathbb{R} \to \mathbb{R}$ *is continuous everywhere in* $\mathbb{R} \setminus \{0,\ z\beta,\ za\}_{z\in\mathbb{Z}}$ *and bounded on all of* $\mathbb{R}$, *and* $f_{per}^{(p)} : \mathbb{R} \to \mathbb{R}$ *is piecewise continuous on* $\mathbb{R}$ *and bounded on* $\mathbb{R} \setminus \{0,\ z\beta,\ za\}_{z\in\mathbb{Z}}$. *Namely,* $f_{per}^{(p)} : \mathbb{R} \to \mathbb{R}$ *may be unbounded at singleton points* $\{0,\ z\beta,\ za\}_{z\in\mathbb{Z}}$.

Then c_j *satisfies*

$$|c_j| \ \leq \ \frac{\beta^{p-1}}{(\pi|j|)^{p}} \left(\begin{array}{l} \beta \| f^{(p)} \|_{\mathrm{L}^\infty(\Omega_2 \setminus \{0,\ \pm a,\ \pm\beta)} \\ \\ +5\| f^{(p-1)} \|_{\mathrm{L}^\infty(\mathbb{R})} \end{array} \right).$$

Hence as $j \to \infty$,

$$|c_j| \ = \ \mathcal{O}\big(\ |j|^{-p} \ \big).$$

For a proof, see Appendix A.2.

Remarks on Theorem 2.2

If $f^{(p-1)}$ is bounded at the singleton points $\{\pm a\}$ in the statement of **(i)**, then the same result holds, except removing references to a from the right hand side of the inequality. Similarly, in the context of **(ii)**, when $f^{(p-1)}$ is smooth at a, and $f^{(p)}$ is bounded at a, the identical inequality holds, except removing all influence by a on the right hand side.

It is noteworthy that the results of Theorem 2.2 generalize in the intuitive way to a function f whose derivatives have singularities of the forms of those in **(i)** and **(ii)** at any finite number of points $\{a_i\}_{i=1}^{p} \subset (0, \beta)$.

The results in Theorem 2.2 can attain better decay estimates for Fourier sine series coefficients than those due to Lemma 2.1. The example of $f(x) := \sin(2\pi x) \,|\sin(2\pi x)| \in \mathrm{C}^1(\mathbb{R}) \setminus \mathrm{C}^2(\mathbb{R})$ in Section 2.2.2 will illustrate this point.

2.2.2 Fourier Sine Series for $f : \Omega \to \mathbb{R}$

Armed with the theory contained in Section 2.2.1, one can examine the decay of Fourier sine series coefficients. For a function $f : \Omega \to \mathbb{R}$, define its odd extension $f_{odd} : \Omega_2 \to \mathbb{R}$ by

$$f_{odd}(x) := \begin{cases} f(x), & x \in \Omega; \\ -f(-x), & x \in [-\beta, 0). \end{cases}$$

Now the full Fourier series of f_{odd} is identical to the Fourier sine series of f in the sense that

$$a_0 + \sum_{j\in\mathbb{N}} \left(a_j \cos\left(\frac{j\pi x}{\beta}\right) + b_j \sin\left(\frac{j\pi x}{\beta}\right)\right) \;=\; \sum_{j\in\mathbb{N}} \widetilde{s}_j \phi_j(x),$$

where

$$\begin{aligned}
\widetilde{s}_j &:= \langle f, \phi_j\rangle_{\Omega}\,, \\
a_0 &:= \frac{1}{2\beta}\langle f_{odd}, 1\rangle_{\Omega_2}\,, \\
a_j &:= \frac{1}{\beta}\left\langle f_{odd}, \cos\left(\frac{j\pi x}{\beta}\right)\right\rangle_{\Omega_2}, \\
b_j &:= \frac{1}{\beta}\left\langle f_{odd}, \sin\left(\frac{j\pi x}{\beta}\right)\right\rangle_{\Omega_2}
\end{aligned}$$

[10, p. 555]. Observe that if $c_j := \frac{1}{2\beta}\left\langle f_{odd}, \overline{\Phi_j}\right\rangle_{\Omega_2}$ is the jth complex Fourier series coefficient of f_{odd}, then

$$\begin{aligned}
a_j - ib_j &= \frac{1}{\beta}\left\langle f_{odd}, \cos\left(\frac{j\pi x}{\beta}\right) + i\sin\left(\frac{j\pi x}{\beta}\right)\right\rangle_{\Omega_2} \\
&= \frac{1}{\beta}\left\langle f_{odd}, \overline{\Phi_j(x)}\right\rangle_{\Omega_2} \\
&= 2c_j.
\end{aligned}$$

Yet notice that for all $j \in \mathbb{N}\cup\{0\}$, $a_j = 0$, and hence

$$b_j \;=\; 2ic_j.$$

Thus

$$\sum_{j\in\mathbb{N}} b_j \sin\left(\frac{j\pi x}{\beta}\right) = \sum_{j\in\mathbb{N}} \widetilde{s}_j \phi_j(x) = \sum_{j\in\mathbb{N}} \widetilde{s}_j \sqrt{\frac{2}{\beta}}\ \sin\left(\frac{j\pi x}{\beta}\right)$$

implies that

$$\widetilde{s}_j = \sqrt{\frac{\beta}{2}}\ b_j = \sqrt{2\beta}\ ic_j.$$

The observation

$$|\widetilde{s}_j| = \sqrt{2\beta}|c_j| \tag{2.13}$$

is critical for estimating the decay of Fourier sine series coefficients. Lemma 2.1 and Theorem 2.2 provide estimates on the decay of $|\widetilde{s}_j|$ by replacing f with $f_{odd} : \Omega_2 \to \mathbb{R}$ in the statements of Lemma 2.1 and Theorem 2.2, respectively. The example below makes use of this idea.

Remark: $f : \Omega \to \mathbb{R}$ Must Satisfy Zero Dirichlet Conditions

Notice that in order to estimate the Fourier sine series coefficients of $f : \Omega \to \mathbb{R}$ by using Theorems 2.1 or 2.2, f must satisfy $f|_{\partial\Omega} = 0$. Otherwise the periodic extension of $f_{odd} : \Omega_2 \to \mathbb{R}$ is discontinuous at even and/or odd integer multiples of β.

Example: $f(x) := \sin(2\pi x)|\sin(2\pi x)| \in \mathrm{C}^1(\Omega_2) \setminus \mathrm{C}^2(\Omega_2)$

An example will illustrate the potential for differences between estimates for $\widetilde{s}_j := \langle f, \phi_j \rangle_\Omega$ due to Lemma 2.1 and Theorem 2.2. Estimates due to Lemma 2.1 can be suboptimal if the periodic extension of $f_{odd}^{(p)} : \Omega_2 \to \mathbb{R}$, the odd extension of $f^{(p)} : \Omega \to \mathbb{R}$, is not bounded at singleton points in Ω_2.

Consider $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) := \sin(2\pi x)|\sin(2\pi x)|$ with $\beta := 1$. The function f is odd and has period 1 so that $(f_{odd})_{per} = (f)_{per} = f$. Define $g(x) := \sin(2\pi x)$.

Consider even j not divisible by 4 (otherwise, $\widetilde{s}_j = 0$). Using integration by parts, a tedious calculation reveals that

$$\left\langle f', \overline{\Phi_j} \right\rangle_{\Omega_2} \;=\; \frac{1}{i\pi j} \left\langle w, \overline{\Phi_j} \right\rangle_{\Omega_2},$$

where $w(x) := (2\ \mathrm{sgn} \circ g)(x)\left[g'(x)^2 + g(x)g''(x)\right]$. By this and Theorem 2.1,

$$c_j := \frac{1}{2}\left\langle f, \overline{\Phi_j} \right\rangle_{\Omega_2} = \frac{1}{2}\left(\frac{1}{i\pi j}\right)\left\langle f', \overline{\Phi_j} \right\rangle_{\Omega_2} = \frac{1}{2}\left(\frac{1}{i\pi j}\right)^2 \left\langle w, \overline{\Phi_j} \right\rangle_{\Omega_2}.$$

Within the contexts of Lemma 2.1 and Theorem 2.2, $w : \mathbb{R} \to \mathbb{R}$ plays the role that $f'' : \mathbb{R} \to \mathbb{R}$ *would* play *were* it to exist. Hence the result of Lemma 2.1 holds by replacing $f^{(p-1)} = f''$ everywhere by w. Namely, $|\widetilde{s}_j| \leq 16\sqrt{2}/j^2$, so that as $j \to \infty$,

$$|\widetilde{s}_j| = \mathcal{O}\big(j^{-2}\big). \tag{2.14}$$

Defining the derivative of the sign function by

$$\operatorname{sgn}'(x) \;:=\; \begin{cases} 0, & x \neq 0; \\ +\infty, & x = 0, \end{cases}$$

one observes that $w' : \mathbb{R} \to \mathbb{R}$ exists but is unbounded on $\{0,\ \pm z/2,\ \pm z\beta\}_{z\in\mathbb{Z}}$. Because w' violates the requirement in Lemma 2.1 that the highest order derivative be bounded, the highest order of decay that Lemma 2.1 predicts for $|\widetilde{s}_j|$ is $\mathcal{O}\left(j^{-2}\right)$.

Yet Theorem 2.2 *does* predict $\mathcal{O}(j^{-3})$ decay by allowing the highest order derivative to have unbounded singularities. The result of Theorem 2.2 **(ii)** holds by replacing $f^{(3)} = f^{(p)}$ by w', so that $|\widetilde{s}_j| \leq 16\sqrt{2}\left(4\pi + 5\right)/(\pi j^3)$, and as $j \to \infty$

$$|\widetilde{s}_j| \;=\; \mathcal{O}\big(j^{-3}\big). \tag{2.15}$$

Figure 2.1 shows that (2.15) accurately estimates the decay rate of $|\widetilde{s}_j|$, while (2.14) is suboptimal (and, hence, so is any bound due to Lemma 2.1).

General Remarks on Fourier Sine Series Coefficients

In general, when $f : \Omega \to \mathbb{R}$ satisfies the Dirichlet boundary conditions $f|_{\partial\Omega} = 0$, as the periodic extension of $f_{odd} : \Omega_2 \to \mathbb{R}$ becomes smoother, assuming that the higher order derivatives also satisfy Dirichlet conditions, the order of the asymptotic decay in j of the coefficients $\widetilde{s}_j := \langle f, \phi_j \rangle_\Omega$ becomes larger. Figures 2.2, 2.3 and 2.4 illustrate this trend.

2.3 Smoothness of Solutions to the Continuous Heat Equation

In the context of IBVP (2.6),

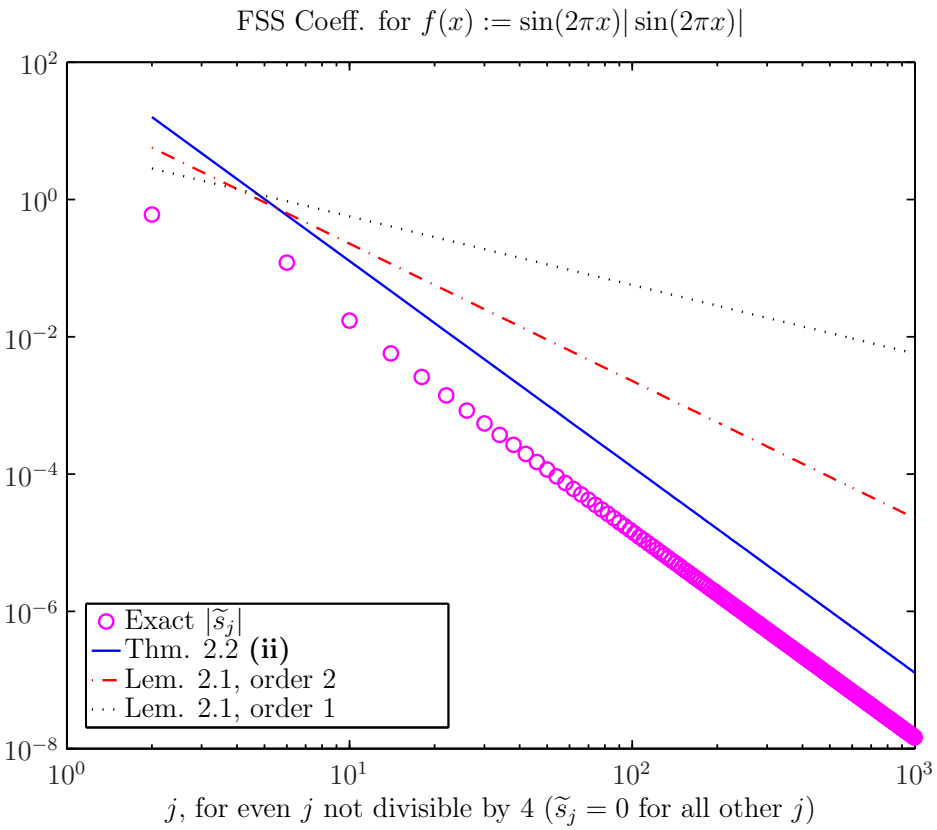


Figure 2.1 : Plotted are magnitudes of Fourier sine series coefficients, $\widetilde{s}_j$, for $f(x) := \sin(2\pi x)|\sin(2\pi x)| \in \mathrm{C}^1(\Omega_2) \setminus \mathrm{C}^2(\Omega_2)$, whose periodic extension is $\mathrm{C}^1 \setminus \mathrm{C}^2(\mathbb{R})$, yet whose coefficeints $|\widetilde{s}_j|$ decay like $\mathcal{O}\left(j^{-3}\right)$. The enhanced error bound of Theorem 2.2 **(ii)** correctly predicts the decay rate of $|\widetilde{s}_j|$, yet Lemma 2.1 predicts only $\mathcal{O}\left(j^{-2}\right)$ decay. Also shown for comparison is the $\mathcal{O}\left(j^{-1}\right)$ estimate from Lemma 2.1.

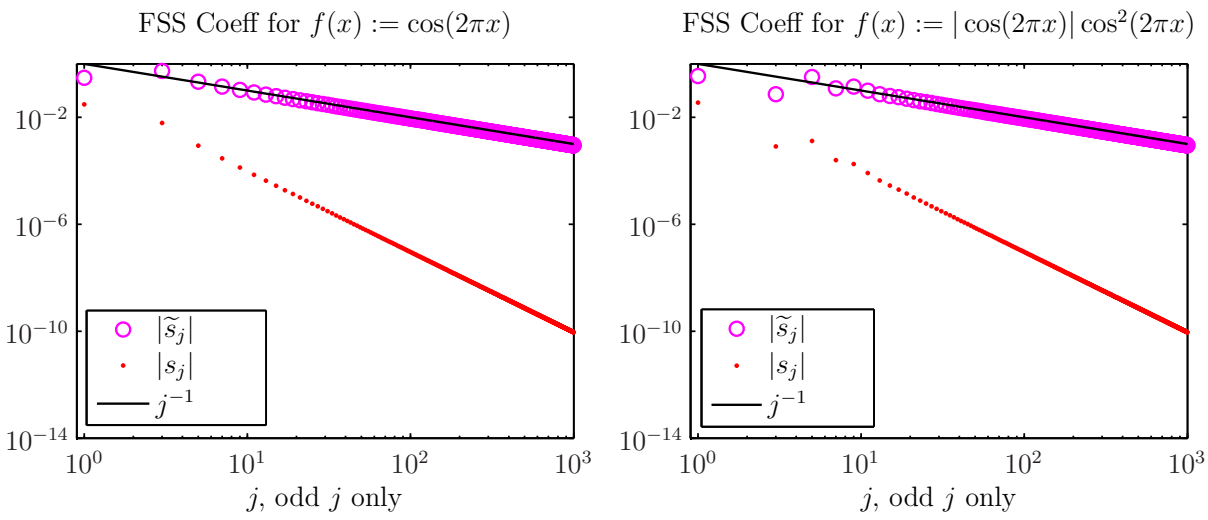


Figure 2.2 : *Left:* $f : \Omega \to \mathbb{R}$ may be smooth, while the periodic extension of $f_{odd} : \Omega_2 \to \mathbb{R}$ is not smooth. For $f(x) := \cos(2\pi x) \in \mathrm{C}^\infty(\mathbb{R})$, both plotted versus j are $|\widetilde{s}_j| := |\langle f, \phi_j \rangle_\Omega|$ (magenta circles) and the upper bound of (2.19) for $\max_{t \geq 0} |s_j(t)|$ (red dots). Because $f|_{\partial\Omega} \neq 0$, the periodic extension of f_{odd} is discontinuous at integers. Asymptotically, $|\widetilde{s}_j| = \mathcal{O}(j^{-1})$.
Right: For $f(x) := |\cos(2\pi x)| \cos^2(2\pi x) \in \mathrm{C}^2(\Omega) \backslash \mathrm{C}^3(\Omega)$, both plotted versus j are $|\widetilde{s}_j|$ (magenta circles) and the upper bound of (2.19) for $\max_{t \geq 0} |s_j(t)|$ (red dots). Because $f|_{\partial\Omega} \neq 0$, the periodic extension of $f_{odd} : \Omega_2 \to \mathbb{R}$ is not $\mathrm{C}^0(\mathbb{R})$. Asymptotically, $|\widetilde{s}_j| = \mathcal{O}(j^{-1})$.

$$
\begin{aligned}
w_t &= \mathrm{L}w + f \quad on \quad \Omega \times [0, T], \\
w|_{\Omega \times \{0\}} &= w_0, \\
w|_{\partial\Omega \times [0,T]} &= 0, \\
w|_{\Omega \times (-\infty, 0)} &= 0,
\end{aligned}
$$

the smoothness of the solution $w : \Omega \times \mathbb{R} \to \mathbb{R}$ at non-negative times is determined by the smoothness of $f : \Omega \times \mathbb{R} \to \mathbb{R}$ at non-negative times and of $w_0 : \Omega \to \mathbb{R}$. (The discussion following Theorem 2.3 elaborates rigorously.)

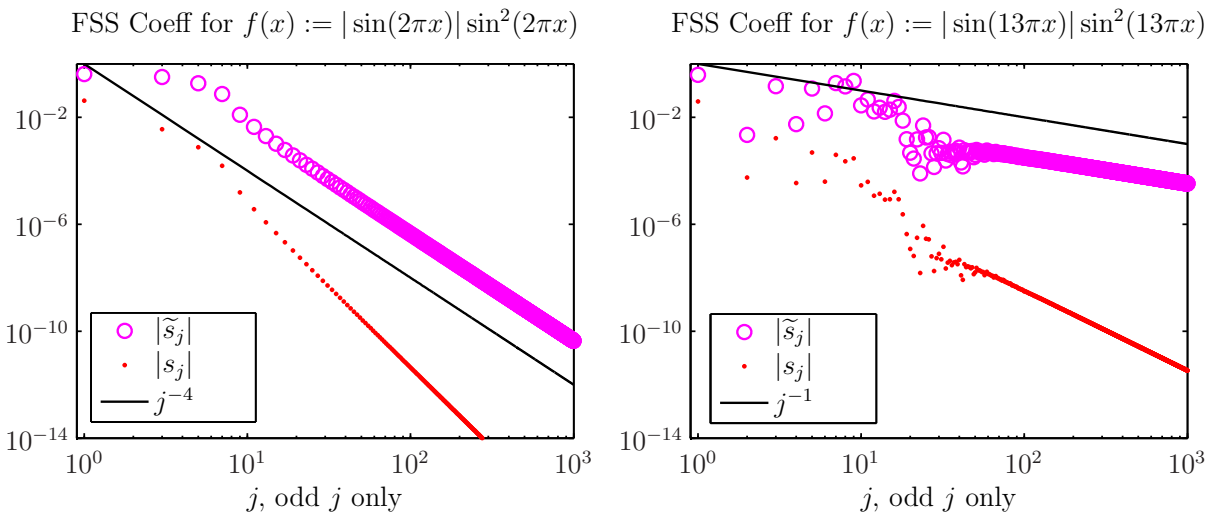


Figure 2.3 : *Left:* For $f(x) := |\sin(2\pi x)|\sin^2(2\pi x) \in \mathrm{C}^2(\Omega) \setminus \mathrm{C}^3(\Omega)$, both plotted versus j are $|\widetilde{s}_j| := |\langle f, \phi_j \rangle_\Omega|$ and the upper bound of (2.19) for $\max_{t\geq 0} |s_j(t)|$. Asymptotically, $|\widetilde{s}_j| = \mathcal{O}(j^{-4})$.
Right: For $f(x) := |\sin(13\pi x)|\sin^2(13\pi x) \in \mathrm{C}^2(\Omega) \setminus \mathrm{C}^3(\Omega)$, both plotted versus j are $|\widetilde{s}_j| := |\langle f, \phi_j \rangle_\Omega|$ and the upper bound of (2.19) for $\max_{t\geq 0} |s_j(t)|$. Asymptotically, $|\widetilde{s}_j| = \mathcal{O}(j^{-1})$; $|\widetilde{s}_j|$ does not achieve the higher order of decay attained in the left plot because $f|_{\partial\Omega} \neq 0$.

Recall that the exact form of w that solves (2.6) is given in (2.7) and (2.8) by

$$w(x,t) \quad = \quad \sum_{j\in\mathbb{N}} s_j(t)\phi_j(x)$$

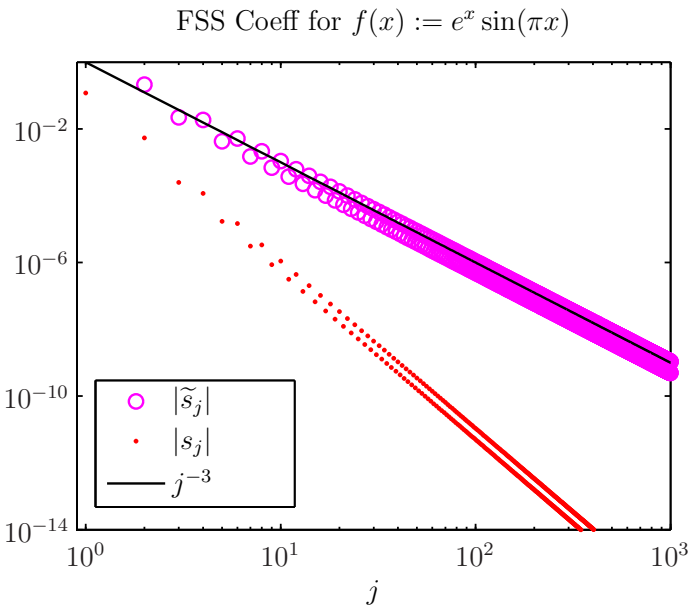


Figure 2.4 : For $f(x) := e^x \sin(\pi x) \in \mathrm{C}^\infty(\mathbb{R})$, both plotted versus j are $|\widetilde{s}_j| := |\langle f, \phi_j \rangle_\Omega| = \mathcal{O}(j^{-3})$ (as $j \to \infty$) and the upper bound of (2.19) for $\max_{t \geq 0} |s_j(t)|$. The second derivative of the periodic extension for $f_{odd} : \Omega_2 \to \mathbb{R}$ is not continuous — though *piecewise* continuous.

for all $t \geq 0$, where for all $t \geq 0$,

$$\begin{aligned} s_j(t) &:= A_j(t) + B_j(t); \\ A_j(t) &:= d_j e^{\lambda_j t} \\ \text{with} \quad d_j &:= \langle w_0, \phi_j \rangle_\Omega ; \\ \text{and} \quad B_j(t) &:= \int_0^t e^{\lambda_j (t-r)} \widetilde{s}_j(r) dr \\ \text{with} \quad \widetilde{s}_j(r) &:= \langle f(\cdot, r), \phi_j \rangle_\Omega . \end{aligned}$$

At a fixed time $t \geq 0$, notice that the contribution made by the basis function $\phi_j : \Omega \to \mathbb{R}$ to w is $s_j(t)$. Observe that at time $t \geq 0$, w is smooth as a function of x

if all non-smooth basis functions contribute insignificantly to w, i.e., if $|s_j(t)|$ is small for all large j. Theorem 2.3 provides estimates on $\{|s_j(t)|\}_{j\in\mathbb{N}}$ for $t \geq 0$.

To facilitate the derivation of a bound on $|s_j(t)|$ for $t \geq 0$, I define a new class of functions.

*Definition 2.1 (**Function** $g_r : \Omega \to \mathbb{R}$)* For fixed $r \in [0,t]$, define $g_r : \Omega \to \mathbb{R}$ by $g_r(x) := f(x,r)$. Whenever it exists, define $p_r \in \mathbb{N}$ to be the maximum parameter with which $g_{r,odd} : \Omega_2 \to \mathbb{R}$, the odd extension of g_r, satisfies the conditions of Theorem 2.1. Denote $g^{(i)}_{r,odd} : \Omega_2 \to \mathbb{R}$ to be the ith derivative of $g_{r,odd}$.

*Theorem 2.3 (**Smoothness of Solutions to the Heat Equation**)*

Fix some arbitrary $t > 0$ at which the decay rate of $|s_j(t)|$ in j is to be estimated. Given the preceding context in Section 2.3, suppose that the Fourier sine series expansion for w converges at all finite times $r \in [0,t]$ and that $\{d_j\}_{j\in\mathbb{N}}$ is a bounded set.

Assume that there exists some p such that $w_{0,odd} : \Omega_2 \to \mathbb{R}$, the odd extension of $w_0 : \Omega \to \mathbb{R}$, satisfies the hypotheses of Theorem 2.1.

Suppose that p_r (see Definition 2.1) exists for all $r \in [0,t]$ and that there exists some $m \in \mathbb{N}$ such that

$$\min_{r\in[0,t]} p_r = m \geq 2.$$

Then as $j \to \infty$,

$$|s_j(t)| \;\leq\; \mathcal{O}\big(j^{-(p-1)}e^{\lambda_j t}\big) + \mathcal{O}\big(j^{-(m+1)}(1-e^{\lambda_j t})\big). \tag{2.16}$$

When $w_0 = 0$, (2.16) *simplifies to*

$$|s_j(t)| \;\leq\; \mathcal{O}\big(j^{-(m+1)}(1-e^{\lambda_j t})\big)$$

as $j \to \infty$.

Proof

To bound $|A_j(t)|$, observe that by (2.13) and applying Theorem 2.1 to $w_{0,odd} : \Omega_2 \to \mathbb{R}$,

$$|A_j(t)| = |d_j e^{\lambda_j t}| \leq \sqrt{2\beta}\; M \left(\frac{\beta}{\pi j}\right)^{p-1} e^{\lambda_j t},$$

where $M := \|w_{0,odd}^{(p-1)}\|_{\mathrm{L}^\infty(\Omega_2)} < \infty$.

For the purpose of bounding $|B_j(t)|$, one must bound $\widetilde{s}_j(r)$ for all $r \in [0,t]$. Consider an arbitrary fixed $r \in [0,t]$ and observe that

$$\widetilde{s}_j(r) \;=\; \int_\Omega f(x,r)\phi_j(x)dx = \langle g_r, \phi_j\rangle_\Omega\,,$$

from which it follows that, because $g_{r,odd} : \Omega_2 \to \mathbb{R}$ satisfies the conditions of Theo-

rem 2.1 with parameter m, by (2.13) and Lemma 2.1,

$$\max_{r\in[0,t]} |\widetilde{s}_j(r)| \;\leq\; \sqrt{2\beta}\;\widetilde{M}\left(\frac{\beta}{\pi j}\right)^{m-1},$$

where $\widetilde{M} := \max_{r\in[0,t]} \|g^{(m-1)}_{r,odd}\|_{\mathrm{L}^\infty(\Omega_2)} < \infty$. Therefore,

$$\begin{aligned}
|B_j(t)| &= \left|\int_0^t e^{\lambda_j(t-r)}\widetilde{s}_j(r)dr\right| \\
&\leq \sqrt{2\beta}\;\widetilde{M}\left(\frac{\beta}{\pi j}\right)^{m-1}\int_0^t e^{\lambda_j(t-r)}dr \\
&= \sqrt{2\beta}\;\widetilde{M}\left(\frac{\beta}{\pi j}\right)^{m-1}\frac{e^{\lambda_j t}-1}{\lambda_j} \\
&= \sqrt{2\beta}\;\widetilde{M}\left(\frac{\beta}{\pi j}\right)^{m+1}(1-e^{\lambda_j t}).
\end{aligned}$$

The claim follows from $|s_j(t)| \leq |A_j(t)| + |B_j(t)|$. ♦

Using Theorem 2.3 to Estimate the Smoothness of Solutions to Continuous Heat Equations

Observe that a non-zero forcing term virtually immediately drowns out all influence of w_0 on the decay of $s_n(t)$ as $t \to \infty$. In particular, if $\beta \approx 1$, for $j \geq 5$ and $r \geq 10^{-1}$, $e^{\lambda_j r} \ll 1$, so that (2.16) gives

$$|s_j(r)| \;=\; \mathcal{O}\left(j^{-(m+1)}\right) \tag{2.17}$$

as $j \to \infty$, indicating that for all times beyond the period immediately following time 0, the smoothness of $w : \Omega \times \mathbb{R} \to \mathbb{R}$ at non-negative times is determined completely by the smoothness of the periodic extensions of the functions $\{\big[f(\cdot, r)\big]_{odd} : \Omega_2 \to \mathbb{R}\}_{r \in [10^{-1}, t]}$.

For the numerical examples remaining, I consider only $w_0 = 0$. Under this assumption,

$$\max_{r \in (0,t]} |s_j(r)| \quad = \quad \mathcal{O}\big(j^{-(m+1)}\big) \tag{2.18}$$

as $j \to \infty$, implying that for larger m, $\max_{r \in (0,t]} |s_j(r)|$ decays more rapidly with j. Recall that m is large if and only if the periodic extensions of $f_{odd}(\cdot, r) : \Omega \to \mathbb{R}$ for all times $r \in [0, t]$ have several continuous spatial derivatives. In particular, the smoother one requires the periodic extensions of $\{\big[f(\cdot, r)\big]_{odd} : \Omega_2 \to \mathbb{R}\}_{r \in [0,t]}$ to be, the more rapidly the coefficients $\{s_j(t)\}_j$ decay with j, and the higher frequency modes are less prevalent in the solution $w : \Omega \times (-\infty, T] \to \mathbb{R}$ at non-negative times.

Several Examples of Decay Estimates for $\max_{t \geq 0} |s_j(t)|$ with j

Consider IBVP (2.6) with $w_0 := 0$ and forcing term $f(x, t) = f_1(x) f_2(t)$. Then

$$s_j(t) \quad = \quad \widetilde{s}_j \int_0^t e^{\lambda_j (t-r)} f_2(r) dr,$$

where $\widetilde{s}_j := \langle f_1, \phi_j \rangle_\Omega$, which implies that

$$\max_{t \geq 0} |s_j(t)| \leq M \frac{|\widetilde{s}_j|}{|\lambda_j|} \tag{2.19}$$

for $M := \|f_2\|_{\mathrm{L}^\infty(\mathbb{R})}$. Hence the decay in j of the Fourier sine series coefficients of $w : \Omega \times \mathbb{R} \to \mathbb{R}$ is an accelerated version of the decay of the Fourier sine series coefficients of the forcing term. Figures 2.2, 2.3 and 2.4 demonstrate this trend by plotting both the right hand side of (2.19) and $|\widetilde{s}_j|$ versus j.

Final Remarks

One expects that in any system Σ that semi-discretizes the continuous heat equation, the finitely many discrete modes of Σ would in some sense approximate a subset of the infinitely many continuous modes. One also anticipates that, in turn, as $f_{odd} : \Omega_2 \times \mathbb{R} \to \mathbb{R}$ becomes smoother in space, the decay in j of the contribution made to the state variable $\mathbf{x} : \mathbb{R} \to \mathbb{R}^n$ at non-negative times by the discrete mode j would behave similarly to (2.18). Both expectations are true and fleshed out in Sections 2.5 and 2.6. First, though, Section 2.4 introduces the finite difference discretization for IBVP (2.6).

2.4 Semi-Discretization Using Centered Finite Differences

To carry the heat equation into the discrete realm, I consider the semi-discretization of IBVP (2.6) using the centered finite difference approximation. Assume that $f :$

$\Omega \times \mathbb{R} \to \mathbb{R}$, $f|_{\Omega\times(-\infty,0)} = 0$, is a product of the form

$$f(x,t) = f_1(x)f_2(t) \tag{2.20}$$

for some $f_1 : \Omega \to \mathbb{R}$ and some $f_2 : \mathbb{R} \to \mathbb{R}$, $f_2|_{(-\infty,0)} = 0$, so that the forcing term is quickly discretized by the product $\mathbf{bu}$ for $\mathbf{u} : \mathbb{R} \to \mathbb{R}$, $\mathbf{u}|_{(-\infty,0)} = 0$.

Discretize (2.6) in its spatial component using the finite difference grid points given by

$$\begin{aligned} z_j &:= jh \\ \text{with } h &:= \frac{\beta}{n+1} \end{aligned}$$

for $j \in \{0, 1, \ldots, n+1\}$. Notice that

$$\frac{z_j}{\beta} = \frac{j}{n+1}.$$

Define the vector of grid points

$$\mathbf{z} := (z_1, \ldots, z_n)^T. \tag{2.21}$$

The centered finite difference approximation is based upon the observation that

$$w''(z_j) = \frac{1}{h^2}\left(w(z_{j-1}) - 2w(z_j) + w(z_{j+1})\right) + \mathcal{O}(h^2)$$

as $n \to \infty$. From this fact, observe that if $w \in \mathcal{Q}$ (see (2.1)), then $\|\mathbf{A}w(\mathbf{z}) - w''(\mathbf{z})\| = \mathcal{O}(h^{3/2})$ as $n \to \infty$, where

$$\mathbf{A} := \frac{1}{h^2} \begin{pmatrix} -2 & 1 & & & \\ 1 & & & & \\ & & \ddots & & \\ & & & & 1 \\ & & & 1 & -2 \end{pmatrix} \in \mathbb{R}^{n \times n}$$

is the finite difference discrete Laplacian. One can show that

$$\sigma(\mathbf{A}) = \left\{ \frac{2}{h^2} \left(-1 + \cos\left(\frac{j\pi}{n+1} \right) \right) \right\}_{j=1}^{n} \tag{2.22}$$

are the eigenvalues of $\mathbf{A}$ with corresponding orthonormal eigenvectors given by

$$\mathbf{v}_j := \tfrac{1}{c_j} \sin\left(\tfrac{j\pi}{n+1} \begin{bmatrix} 1 \\ \vdots \\ n \end{bmatrix} \right) = \tfrac{1}{c_j} \sin\left(\tfrac{j\pi}{\beta} \mathbf{z} \right), \tag{2.23}$$

where

$$c_j := \sqrt{\textstyle\sum_{m=1}^{n} \sin\left(\frac{mj\pi}{n+1} \right)^2} \tag{2.24}$$

is chosen so that $\|\mathbf{v}_j\| = 1$ [12, p. 8, 16, 276].

Denote the jth eigenvalue of $\mathbf{A} \in \mathbb{R}^{n \times n}$ by $\mu_j^{(n)}$ $(0 > \mu_1^{(n)} > \cdots > \mu_{n-1}^{(n)} > \mu_n^{(n)})$ and take

$$\mathbf{M} := \begin{pmatrix} \mu_1^{(n)} & & \\ & \ddots & \\ & & \mu_n^{(n)} \end{pmatrix} \in \mathbb{R}^{n \times n}.$$

Then

$$\mathbf{A} = \mathbf{V}\mathbf{M}\mathbf{V}^* \tag{2.25}$$

is a unitary diagonalization of $\mathbf{A}$.

Given the form of f in (2.20),

$$f(\mathbf{z}, t) \approx \mathbf{b}\ \mathbf{u}(t) \in \mathbb{R}^n, \tag{2.26}$$

where $\mathbf{b} := f(\mathbf{z})$ and $\mathbf{u}(t) := f_2(t) \in \mathbb{R}$.

In this context, the *semi-discretized* version of the IBVP (2.6) is the initial value problem (IVP)

$$\begin{array}{lcll} \dot{\mathbf{x}} & = & \mathbf{A}\mathbf{x} + \mathbf{b}\mathbf{u} & \text{for } t \geq 0, \\ \mathbf{x} & = & \mathbf{0} & \text{for } t < 0, \\ \mathbf{x}(0) & = & w_0(\mathbf{z}), & \end{array} \tag{2.27}$$

where $\mathbf{x}_j(t) \approx w(z_j, t)$. Notice that the zero boundary conditions are automatically enforced by ignoring the contributions from the boundary nodes z_0 and z_{n+1} in (2.27) (see, e.g., [20, p. 62] for explanation).

2.5 Relating the Semi-Discretized and Continuous Problems: Convergence as n Grows

As the semi-discretized heat equation (2.27) approximates the continuous heat equation (2.6), there is a correspondence between the discrete modes and a finite subset

of the continuous modes. Here I establish this idea, showing that each discrete eigenvalue and eigenvector approximates a continuous eigenvalue and eigenfunction in some sense.

2.5.1 Eigenvalues of the Continuous and Semi-Discretized Problems

*Lemma 2.2 (**Convergence of the Discrete to the Continuous Eigenvalues**)*

For fixed $j \in \mathbb{N}$,

$$\lim_{n\to\infty} \mu_j^{(n)} = \lambda_j.$$

Namely, for n adequately large,

$$0 < \mu_j^{(n)} - \lambda_j < \frac{2(j\pi)^4}{\beta^2(n+1)^2}\left(\cosh(1) - 3/2\right), \tag{2.28}$$

and

$$\mu_j^{(n)} \;=\; \lambda_j + \mathcal{O}\left(n^{-2}\right)$$

as $n \to \infty$.

Proof

Assume that $n \in \mathbb{N}$ is adequately large to satisfy

$$\frac{(j\pi)}{n+1} < 1.$$

From the Taylor series expansion for the cosine function, observe that

$$\begin{aligned}
\mu_j^{(n)} &= \frac{2}{h^2}\left(-1 + \cos\left(\frac{j\pi}{n+1}\right)\right) \\
&= \frac{2\,(n+1)^2}{\beta^2}\left(-1 + 1 - \frac{1}{2!}\left(\frac{j\pi}{n+1}\right)^2 + \frac{1}{4!}\left(\frac{j\pi}{n+1}\right)^4 - \frac{1}{6!}\left(\frac{j\pi}{n+1}\right)^6 + \cdots\right) \\
&= -\left(\frac{j\pi}{\beta}\right)^2 + \frac{2}{\beta^2}\left(\frac{(j\pi)^4}{4!\,(n+1)^2} - \frac{(j\pi)^6}{6!\,(n+1)^4} + \cdots\right) \\
&=: \lambda_j + \frac{2}{\beta^2}\vartheta,
\end{aligned}$$

where $\vartheta := \nu_4 - \nu_6 + \cdots$, and $\nu_k := \frac{(j\pi)^k}{k!\,(n+1)^{k-2}} > 0$. Notice that $\nu_k > \nu_{k+2}$ because

$$\begin{aligned}
\nu_{k+2} &= \frac{(j\pi)^{k+2}}{(k+2)!\,(n+1)^k} \\
&= \frac{(j\pi)^2}{(k+1)(k+2)(n+1)^2}\frac{(j\pi)^k}{k!\,(n+1)^{k-2}} \\
&= \frac{(j\pi)^2}{(k+1)(k+2)(n+1)^2}\nu_k \\
&< \nu_k.
\end{aligned}$$

Thus

$$\vartheta \;:=\; \sum_{k\in\{2,\ 4,\ 6,\ldots\}} (\nu_{2k} - \nu_{2k+2}) > 0,$$

and

$$\mu_j^{(n)} - \lambda_j = \frac{2}{\beta^2}\vartheta > 0,$$

so that $|\mu_j^{(n)} - \lambda_j| = \mu_j^{(n)} - \lambda_j$. Now ϑ can be bounded by observing that

$$\begin{aligned}
\vartheta &= \sum_{z=2}^{\infty}(-1)^z\nu_{2z} \\
&< \sum_{z=2}^{\infty}\nu_{2z} \\
&= (j\pi)^2\sum_{z=2}^{\infty}\frac{1}{(2z)!}\left(\frac{j\pi}{n+1}\right)^{2z-2} \\
&< (j\pi)^2\left(\frac{j\pi}{n+1}\right)^2\sum_{z=2}^{\infty}\frac{1}{(2z)!} \\
&= \frac{(j\pi)^4}{(n+1)^2}\left(\cosh(1) - \frac{3}{2}\right),
\end{aligned}$$

where I have used the fact that

$$\sum_{z=0}^{\infty}\frac{1}{(2z)!} = \cosh(1) \;\;\Rightarrow\;\; \sum_{z=2}^{\infty}\frac{1}{(2z)!} = \cosh(1) - \frac{3}{2}$$

[18]. Hence

$$|\mu_j^{(n)} - \lambda_j| = \mu_j^{(n)} - \lambda_j = \frac{2}{\beta^2}\,\vartheta < \frac{2(j\pi)^4}{\beta^2(n+1)^2}\,(\cosh(1) - 3/2)\,,$$

and $|\mu_j^{(n)} - \lambda_j| = \mathcal{O}\,(n^{-2})$ as $n \to \infty$. ♦

By Lemma 2.2, when the quantity $(j\pi)^4/[\beta^2(n+1)^2]$ is small, the approximation $\mu_j^{(n)} \approx \lambda_j$ is good. Consider the eigenvalue for which this quantity is always the smallest, i.e., the lowest frequency eigenvalue, and set $\beta := 1$ so that

$$0 < \mu_1^{(n)} - \lambda_1 < \frac{2\pi^4}{(n+1)^2}\,(\cosh(1) - 3/2)\,. \tag{2.29}$$

As illustrated in Figure 2.5, $|\mu_1^{(n)} - \lambda_1| < 10^{-1}$ by $n = 10$. As j grows, n must also grow to yield an accurate approximation $\mu_j^{(n)} \approx \lambda_j$. Figure 2.5 shows that n must be considerably larger (roughly 10^3) to yield $|\mu_{10} - \lambda_{10}| \leq 10^{-1}$.

Suppose that the index $j_n \leq n$ grows with n. If j_n is close to n, then $\mu_j^{(n)} = \mathcal{O}(n^2)$ as $n \to \infty$ because

$$\frac{4(n+1)^2}{\beta^2} \geq |\mu_{j_n}^{(n)}| \geq \frac{3(n+1)^2}{\beta^2}. \tag{2.30}$$

This follows from the observation that

$$\begin{aligned}\frac{4(n+1)^2}{\beta^2} = \frac{4}{h^2} &\geq \left|\frac{2}{h^2}\left(-1+\cos\left(\frac{j_n\pi}{n+1}\right)\right)\right| \\ &= \left|\frac{2}{h^2}\left(1+\left|\cos\left(\frac{j_n\pi}{n+1}\right)\right|\right)\right| \\ &> \frac{3}{h^2} = \frac{3(n+1)^2}{\beta^2},\end{aligned}$$

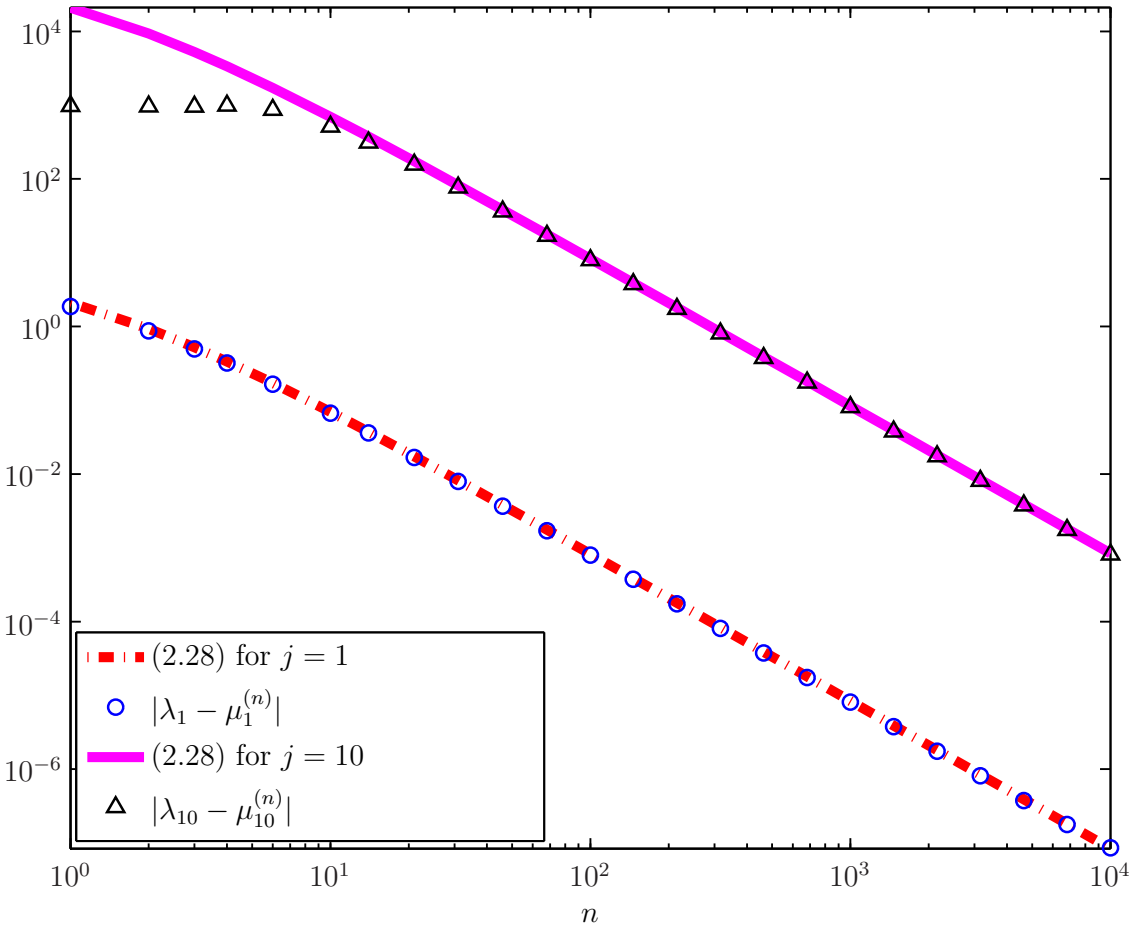


Figure 2.5 : The plot illustrates the convergence of $\mu_j^{(n)} \to \lambda_j$ as $n \to \infty$ for $j \in \{1,\ 10\}$. Plotted are both the right hand side of (2.28) versus n and $|\lambda_j - \mu_j^{(n)}|$ versus n for $j \in \{1,\ 10\}$.

where I have used the observation that given sufficiently large n, for sufficiently large values of j_n, $\left|\cos\left(\frac{j_n\pi}{n+1}\right)\right| > 1/2$.

The convergence $\mu_j^{(n)} \to \lambda_j$ given by Lemma 2.2 does not hold when $j \in \mathbb{N}$ depends on n. Specifically, the highest frequency $\mu_j^{(n)}$ values tend to give poor approximations to corresponding λ_j values. This holds most clearly for $j = n$. The upper bound from Lemma 2.2 correctly estimates that the error $|\mu_n^{(n)} - \lambda_n|$ is significant. Observe

$$\begin{aligned} 0 < \mu_n^{(n)} - \lambda_n &< \frac{2(n\pi)^4}{\beta^2(n+1)^2}\left(\cosh(1) - 3/2\right) \\ &= \mathcal{O}\left(n^2\right) \end{aligned}$$

as $n \to \infty$, reflecting the reality that $\mu_n^{(n)}$ is a remarkably poor approximation to λ_n. Figure 2.6 shows the symbolically calculated relative error $|\lambda_n - \mu_n^{(n)}|/||\lambda_n|$ versus n. The relative error converges to roughly 0.6 as n grows.

Eigenfunctions and Eigenvectors of the Continuous and Semi-Discretized Problems

Corresponding to the convergence $\mu_j^{(n)} \to \lambda_j$, the jth eigenvector of $\mathbf{A}$ also converges to the jth eigenfunction of $\mathrm{L} : \mathcal{Q} \to \mathrm{L}^2(\Omega)$ in some sense. Take

$$\rho_j(x) := \sin\left(\frac{j\pi}{\beta}x\right) = \sqrt{\frac{\beta}{2}}\phi_j(x), \tag{2.31}$$

the jth eigenfunction of L. One observes immediately that $c_j\mathbf{v}_j$ approximates $\rho_j(\mathbf{z})$ exactly on the grid $\mathbf{z}$, yet more can be said. Before doing so in Lemma 2.4, I state an important result from interpolation theory.

*Lemma 2.3 (**Interpolation Error Bound**)*

If the degree n polynomial $p_n : [a,b] \to \mathbb{R}$ interpolates $f \in C^{n+1}[a,b]$ at the distinct points $\{z_j\}_{j=0}^n \subset [a,b]$, then

$$\|f - p_n\|_{\mathrm{L}^\infty[a,b]} \leq \frac{\|f^{(n+1)}\|_{\mathrm{L}^\infty[a,b]}}{(n+1)!} \left(\max_{x\in[a,b]} \prod_{j=0}^{n} |x - z_j| \right). \tag{2.32}$$

(See, e.g., [17, p. 183-184].)

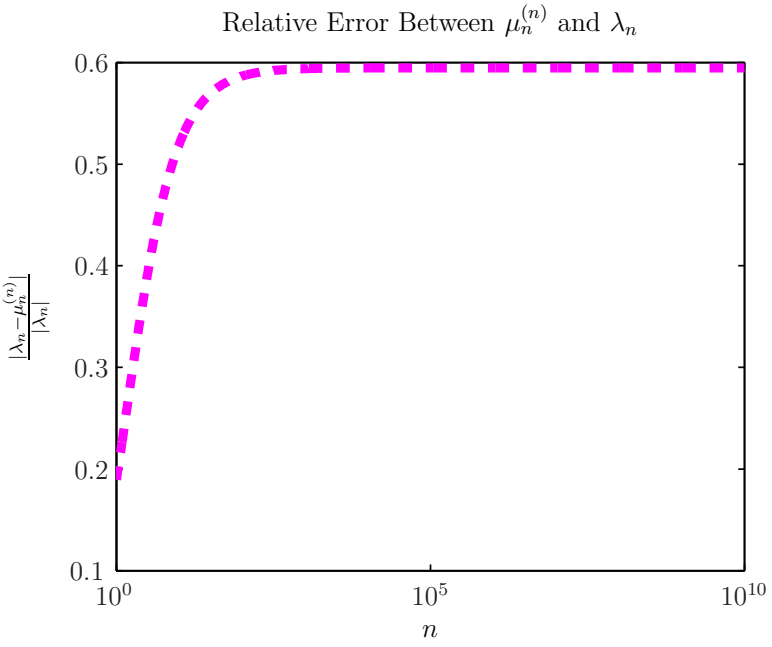


Figure 2.6 : Illustration of the poor relative error between λ_n and $\mu_n^{(n)}$. Here, values of $\mu_n^{(n)}$ and λ_n are symbolically calculated.

*Lemma 2.4 (**Convergence of Discrete Eigenvectors to Eigenfunctions**)*

For fixed $j \in \mathbb{N}$, if $S : \Omega \to \mathbb{R}$ is the piecewise linear interpolant to the points $\{ (\mathbf{z}(i),\ \mathbf{v}_j(i)\ ) \}_{i=1}^{n} \cup \{ (0,0), (\beta,0) \}$, then

$$\|\rho_j - S\|_{\mathrm{L}^2(\Omega)} \leq \frac{\sqrt{\beta}}{2} \left(\frac{\pi j}{n+1} \right)^2 = \mathcal{O}\left(n^{-2}\right) \to 0$$

as $n \to \infty$.

Proof

Define $\Omega_i := [z_{i-1}, z_i] \subset \Omega$. Then $S|_{\Omega_i}$, the restriction of $S : \Omega \to \mathbb{R}$ to Ω_i, interpolates the points $(z_{i-1},\ \rho_j(z_{i-1}))$ and $(z_i,\ \rho_j(z_i))$. Hence, by (2.32),

$$\begin{aligned} \|\rho_j - S\|_{\mathrm{L}^\infty(\Omega_i)} &\leq \frac{\|\rho_j''\|_{\mathrm{L}^\infty(\Omega_i)}}{2} \left(\frac{\beta}{n+1} \right)^2 \\ &\leq \frac{1}{2} \left(\frac{\pi j}{\beta} \right)^2 \left(\frac{\beta}{n+1} \right)^2 \\ &= \frac{1}{2} \left(\frac{\pi j}{n+1} \right)^2, \end{aligned}$$

where I have used Lemma 2.3 and the fact that

$$\rho_j''(x) = -\left(\frac{\pi j}{\beta} \right)^2 \sin\left(\frac{\pi j}{\beta} x \right) \quad \Rightarrow \quad \|\rho_j''\|_{\mathrm{L}^\infty(\Omega)} = \left(\frac{\pi j}{\beta} \right)^2.$$

It follows then that

$$\begin{aligned}
\|\rho_j - S\|^2_{\mathrm{L}^2(\Omega)} &= \int_\Omega (\rho_j(x) - S(x))^2 dx \\
&= \sum_{i=1}^{n+1} \int_{\Omega_i} (\rho_j(x) - S(x))^2 dx \\
&\leq \sum_{i=1}^{n+1} \int_{\Omega_i} \left(\frac{1}{2} \left(\frac{\pi j}{n+1} \right)^2 \right)^2 dx \\
&= (n+1) \left(\frac{\beta}{n+1} \right) \left(\frac{1}{2} \left(\frac{\pi j}{n+1} \right)^2 \right)^2 \\
&= \frac{\beta}{4} \left(\frac{\pi j}{n+1} \right)^4 .
\end{aligned}$$

Consequently,

$$\|\rho_j - S\|_{\mathrm{L}^2(\Omega)} \leq \frac{\sqrt{\beta}}{2} \left(\frac{\pi j}{n+1} \right)^2 = \mathcal{O}\left(n^{-2}\right)$$

as $n \to \infty$. ♦

Observe then from Lemmas 2.2 and 2.4 that mode j of the semi-discrete heat equation approximates mode j of the continuous problem in the sense that both $\mu_j^{(n)} \approx \lambda_j$, and the piecewise linear interpolant determined by the eigenvector $\mathbf{v}_j$ and the finite difference grid $\mathbf{z}$ converges in $\mathrm{L}^2(\Omega)$ to a multiple of the jth eigenfunction of L. For fixed $j \ll n$, this approximation becomes more accurate as n grows. Consequently, the importance of the jth discrete mode for $j \ll n$ to the discrete system should be similar to the importance of its continuous counterpart to the continuous problem, i.e., the contribution should decay significantly as j increases. Section 2.6 rigorously

establishes that hypothesis.

2.6 "Smoothness" in the Discrete Sense

The notion of smoothness in space for a function $w : \Omega \to \mathbb{R}$ can be generalized to a similar notion of *discrete smoothness*. Roughly speaking, a vector $\mathbf{x} \in \mathbb{R}^n$ can be called *discretely smooth* if it has minimal influence from all eigenvectors $\mathbf{v}_j \in \mathbb{R}^n$ of the discrete Laplacian corresponding to the eigenvalues $\mu_j^{(n)}$ of large magnitude. That is,

$$\mathbf{x} = \sum_{j=1}^{n} a_j \mathbf{v}_j, \tag{2.33}$$

where $|a_j|$ is small for all large j.

It follows then that the solution $\mathbf{x} : \mathbb{R} \to \mathbb{R}^n$ to the linear, time-invariant system

$$\begin{array}{lcll} \dot{\mathbf{x}} & = & \mathbf{A}\mathbf{x} + \mathbf{b}\mathbf{u} & \text{for } t \geq 0, \\ \mathbf{x} & = & \mathbf{0} & \text{for } t < 0, \\ \mathbf{x}(0) & = & w_0(\mathbf{z}), & \end{array} \tag{2.34}$$

can be called smooth for non-negative times when $\mathbf{x}(t)$ is discretely smooth for all fixed times $t \geq 0$, where $\mathbf{z} \in \mathbb{R}^n$ is the finite difference grid on Ω. That is, for $t \geq 0$,

$$\mathbf{x}(t) = \sum_{j=1}^{n} a_j(t) \mathbf{v}_j, \tag{2.35}$$

where $\max_{t \geq 0} |a_j(t)|$ decays rapidly as j increases. In such a case, for all fixed $t \geq 0$, there exists some smooth $w^{(t)} : \Omega \to \mathbb{R}$ that is well approximated by $\mathbf{x}(t)$ in the sense that $\mathbf{x}_p(t) \approx w^{(t)}(z_p)$ for all $p \in \{1, \ldots, n\}$.

To measure the influence of the high frequency eigenmodes on the solution $\mathbf{x} : \mathbb{R} \to \mathbb{R}^n$ to IVP (2.34) at non-negative times, one must first develop a sense for how the coefficients in a discrete representation for a time-independent function $g : \Omega \to \mathbb{R}$, with the convergent Fourier sine series

$$g(x) \;=\; \sum_{j \in \mathbb{N}} s_j \phi_j(x),$$

decay. To this end, Section 2.6.1 gives discrete analogues to Theorems 2.1 and 2.2.

Following that discussion, Section 2.6.2 generalizes Theorem 2.3 to the discrete realm by establishing an analogous idea for the solution $\mathbf{x} : \mathbb{R} \to \mathbb{R}^n$ to the linear, time-invariant system (2.34).

2.6.1 Discrete Analogues to Theorems 2.1 and 2.2

Select some function $g : \Omega \to \mathbb{R}$ with a convergent Fourier sine series given by

$$\begin{aligned} g(x) &= \sum_{j \in \mathbb{N}} s_j \phi_j(x), \\ s_j &:= \langle g, \phi_j \rangle_\Omega \,. \end{aligned}$$

Assume that $g_{odd} : \Omega_2 \to \mathbb{R}$ satisfies the conditions of either Lemma 2.1 or Theorem 2.2 for some $p > 2$, so that the Fourier sine series coefficients decay like

$$|s_j| = \mathcal{O}\big(j^{-(p-1)}\big) \tag{2.36}$$

as $j \to \infty$. Define $g_n : \Omega_2 \to \mathbb{R}$ by

$$g_n(x) := \sum_{j=1}^{n} s_j \phi_j(x). \tag{2.37}$$

Then g_n is the best approximation in the $\mathrm{L}^2(\Omega)$ norm to g from $\mathrm{span}\{\phi_1, \cdots, \phi_n\}$ [10, p. 142].

Lemma 2.5 ($g_n \to g$)

Suppose that $g : \Omega \to \mathbb{R}$ has the convergent Fourier sine series

$$g(x) \;=\; \sum_{j \in \mathbb{N}} s_j \phi_j(x),$$

where $s_j := \langle g, \phi_j \rangle_\Omega$, and $g_{odd} : \Omega_2 \to \mathbb{R}$ satisfies the conditions of either Lemma 2.1 or Theorem 2.2 for some $p > 2$, so that there exists some p satisfying

$$|s_j| = \mathcal{O}\big(j^{-(p-1)}\big)$$

as $j \to \infty$. Then

$$g_n \;\to\; g$$

(see (2.37)) as $n \to \infty$ in both $\|\cdot\|_{\mathrm{L}^2(\Omega)}$ and $\|\cdot\|_{\mathrm{L}^\infty(\Omega)}$.

Proof

I begin by showing that $\|g(x) - g_n(x)\|_{\mathrm{L}^2(\Omega)} \to 0$. Observe that $\|\phi_n\|_{\mathrm{L}^2(\Omega)} = 1$, and hence

$$\begin{aligned}
\|g(x) - g_n(x)\|_{\mathrm{L}^2(\Omega)} &= \left\| \sum_{j=n+1}^{\infty} s_j \phi_j(x) \right\|_{\mathrm{L}^2(\Omega)} \\
&\leq \sum_{j=n+1}^{\infty} |s_j| \\
&= \mathcal{O}\left(\sum_{j=n+1}^{\infty} j^{-(p-1)} \right) \\
&\to 0
\end{aligned}$$

as $n \to \infty$ because $\sum_{j=1}^{\infty} j^{-(p-1)} < \infty$ for $p > 2$.

Using the bound $\|\phi_j\|_{\mathrm{L}^\infty(\Omega)} \leq \sqrt{\frac{2}{\beta}}$, $\|g(x) - g_n(x)\|_{\mathrm{L}^\infty(\Omega)} \to 0$ follows by identical reasoning. ♦

Recall the function $\rho_j : \Omega \to \mathbb{R}$ from (2.31) given by

$$\rho_j(x) := \sqrt{\frac{\beta}{2}}\, \phi_j(x)$$

and define

$$\mathbf{q}_j := c_j \mathbf{v}_j \in \mathbb{R}^n,$$

where ϕ_j, $\mathbf{v}_j$ and c_j are given in (2.3), (2.23) and (2.24) respectively. Then

$$\mathbf{q}_j \;=\; \rho_j(\mathbf{z})$$

approximates the jth eigenfunction of $\mathrm{L} : \mathcal{Q} \to \mathrm{L}^2(\Omega)$ (see (2.2)) exactly on the finite difference grid $\mathbf{z} \in \mathbb{R}^n$.

By Lemma 2.5, for arbitrary $\epsilon > 0$, there exists $n_\epsilon \in \mathbb{N}$ such that all $n \geq n_\epsilon$ satisfy $\|g - g_n\|_{\mathrm{L}^\infty(\Omega)} < \epsilon$. Consider then the approximate discretization of g given by

$$\begin{aligned} g(\mathbf{z}) &\approx g_n(\mathbf{z}) \\ &= \textstyle\sum_{j=1}^n s_j \phi_j(\mathbf{z}) \\ &= \textstyle\sum_{j=1}^n \sqrt{\frac{2}{\beta}}\; s_j c_j \mathbf{v}_j \\ &= \textstyle\sum_{j=1}^n a_j \mathbf{v}_j, \end{aligned} \tag{2.38}$$

where $a_j := \sqrt{\frac{2}{\beta}}\; s_j c_j$, and I have used the fact that

$$\sqrt{\frac{\beta}{2}}\phi_j(\mathbf{z}) = \rho_j(\mathbf{z}) = \mathbf{q}_j = c_j \mathbf{v}_j.$$

Combining the bounds for $|a_j|$ from Theorems 2.1 and 2.2 with (2.38) yields approximate descriptions for coefficient decay in the expansion of $g(\mathbf{z})$ using the eigenvector basis $\{\mathbf{v}_j\}_{j=1}^n$ of $\mathbf{A}$. These discrete analogues for Theorems 2.1 and 2.2 apply immediately in estimating the magnitudes of the time-dependent coefficients $\{a_j(t)\}_{j=1}^n$ from (2.35).

2.6.2 Smoothness of the Solution $\mathbf{x} : \mathbb{R} \to \mathbb{R}^n$

Assume that the Fourier series expansion for the solution (2.7)

$$w(x,t) = \sum_{j \in \mathbb{N}} s_j(t)\phi_j(x) \tag{2.39}$$

converges for all times and all $x \in \Omega$. Assume also that $[w(\cdot,t)]_{odd} : \Omega_2 \to \mathbb{R}$ satisfies the conditions of either Lemma 2.1 or Theorem 2.2 for some $p > 2$ for all $t \geq 0$. By Lemma 2.5, for arbitrary $\epsilon > 0$, for any fixed $t \geq 0$, there exists $n_t \in \mathbb{N}$ such that $\|w(\mathbf{z},t) - w_{n_t}(\mathbf{z},t)\| < \epsilon$, where

$$w_{n_t}(x,t) \;\; := \;\; \sum_{j=1}^{n_t} s_j(t)\phi_j(x).$$

Consider some finite collection of positive times $\{t_p\}_{p=1}^{q}$ defined by

$$t_p := ph,$$

where $h > 0$ is a time-step, potentially to be used with forward Euler. Notice that there exists some $n < \infty$ such that

$$\max_{\{t_p : 1 \leq p \leq q\}} n_{t_p} = n.$$

Then by (2.38), for any t_p,

$$\begin{aligned} \mathbf{x}(t_p) &\approx w(\mathbf{z}, t_p) \\ &\approx \textstyle\sum_{j=1}^{n} a_j(t_p)\ \mathbf{v}_j, \end{aligned} \tag{2.40}$$

where

$$a_j(t) := \sqrt{\frac{2}{\beta}}\ c_j\ s_j(t), \tag{2.41}$$

and $s_j(t)$ is given in (2.39).

Now Theorem 2.3 estimates the decay rate of $|s_j(t)|$ in j, allowing one to estimate the decay rate of $|a_j(t_p)|$ in j, and in turn to estimate the smoothness of $\mathbf{x} : \{t_p\}_{p=1}^{q} \to \mathbb{R}^n$. In particular, if the assumptions of Theorem 2.3 hold and $w_0 = 0$, then

$$|a_j(t_p)| = \mathcal{O}\left(j^{-(m+1)}\right) \tag{2.42}$$

as $j \to \infty$. The smoother in space one requires the periodic extensions of $f_{odd} : \Omega_2 \times \mathbb{R} \to \mathbb{R}$ to be at all times $\{t_p\}_{p=1}^{q}$, the larger the value of the parameter m. In turn, $|a_j(t_p)|$ decays more rapidly as $j \to \infty$, and thus $\mathbf{x} : \{t_p\}_{p=1}^{q} \to \mathbb{R}^n$ becomes smoother.

Consequently, for a larger value of m, higher frequency modes contribute less significantly to $\mathbf{x} : \{t_p\}_{p=1}^{q} \to \mathbb{R}^n$, and thus their omission from Σ results in a negligible loss of accuracy in the state variable. Section 2.7 establishes this idea rigorously.

2.7 Dual-Stage Dimension Reduction of the Semi-Discretized Problem

Equation (2.42) has important implications for the stable time-step for the forward Euler scheme. For spatially smooth $f_{odd} : \Omega_2 \times \mathbb{R} \to \mathbb{R}$, all modes corresponding to large j are unimportant in contributing to the state variable $\mathbf{x}$, and ignoring them would therefore yield little loss of accuracy, while increasing the system's stable time-step by orders of magnitude. Modal filtering acting on Σ would accomplish just that, but its cost renders it unjustifiable for adequately large n.

Yet modal reduction is more palatable when applied to a sufficiently small reduced-order model obtained from the original model by using moment matching. Moreover, as $\widehat{\Sigma} \approx \Sigma$, the high frequency modes of $\widehat{\Sigma}$ are unimportant, just as the high frequency modes of Σ are unimportant. It follows that filtering them from $\widehat{\Sigma}$ leads to a negligible loss of accuracy. Section 2.7 makes this clear and develops a bound on the absolute value of the difference between the transfer function of $\widehat{\Sigma}$ and its modally filtered counterpart to yield an intelligent choice for the truncation mode.

2.7.1 Moment Matching

Take the unitary diagonalization $\mathbf{A} = \mathbf{VMV}^*$ of (2.25). Refer to the system Σ as the semi-discretized heat equation of the form (2.34) given by

$$\Sigma := \left(\begin{array}{lcll} \dot{\mathbf{x}} & = & \mathbf{A}\mathbf{x} + \mathbf{b}\mathbf{u} & \text{for } t \geq 0 \\ \mathbf{x} & = & \mathbf{0} & \text{for } t < 0 \\ \mathbf{x}(0) & = & \mathbf{0} & \\ \mathbf{y} & = & \mathbf{c}\mathbf{x} & \text{for all } t \end{array} \right),$$

with the corresponding transfer function

$$\mathcal{H}(s) = \mathbf{c}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}.$$

Define $\mathbf{s} \in \mathbb{C}^n$ and $\mathbf{r} \in \mathbb{C}^{1\times n}$ by

$$\begin{array}{rcl} \mathbf{s} & := & \mathbf{V}^*\mathbf{b}, \\ \mathbf{r} & := & \mathbf{c}\mathbf{V}, \end{array} \tag{2.43}$$

so that

$$\begin{array}{rcl} \mathbf{b} & = & \displaystyle\sum_{j=1}^{n} s_j \mathbf{v}_j \in \mathbb{C}^n, \\ \mathbf{c} & = & \displaystyle\sum_{j=1}^{n} r_j \mathbf{v}_j^* \in \mathbb{C}^{1\times n}. \end{array}$$

Realize that for most linear, time-invariant systems Σ with $\mathbf{A} \in \mathbb{R}^{n\times n}$ that is Hermitian without any eigenvalues close to 0, the transfer function $|\mathcal{H}|$ is relatively small when evaluated on the imaginary axis because $\sigma(\mathbf{A}) \subset \mathbb{R}$, and for any $\omega \in \mathbb{R}$, $\|(i\omega\mathbf{I} - \mathbf{A})^{-1}\| = \max_{1\leq j\leq n} 1/(\sqrt{\omega^2 + \lambda_j^2})$ is relatively small. If it also holds that

$|(\mathcal{L}\mathbf{u})(i\omega)|$ is maximized at frequencies near 0 while insignificant at frequencies far from 0, then the Laplace transform of the output

$$(\mathcal{L}\mathbf{y})(i\omega) = \mathcal{H}(i\omega)(\mathcal{L}\mathbf{u})(i\omega)$$

is likewise most significant at frequencies ω near 0 and less significant elsewhere.

Recall from the discussion surrounding (1.3) that to match the output of Σ with that of a reduced-order model $\widehat{\Sigma}$ by moment matching, one matches $\mathcal{H}(i\omega) \approx \widehat{\mathcal{H}}(i\omega)$ at frequencies $i\omega$ where $|(\mathcal{L}\mathbf{y})(i\omega)|$ is largest. Hence, for systems such that $|(\mathcal{L}\mathbf{y})(i\omega)|$ is largest for frequencies near 0, matching moments of $\mathcal{H}$ for the series expansion about 0 is ideal. One accomplishes this by requiring that

$$\widehat{\mathbf{c}}\widehat{\mathbf{A}}^{-j}\widehat{\mathbf{b}} = \mathbf{c}\mathbf{A}^{-j}\mathbf{b} \tag{2.44}$$

for $j = 1, \ldots, k$. Recall from Theorem 1.1 that inverted Arnoldi, which generates a basis for $\mathcal{K}_k(\mathbf{A}^{-1}, \mathbf{b})$, attains (2.44). Nonetheless, moment matching via inverted Arnoldi assumes that n is small enough to justify computations with $\mathbf{A}^{-1}$. Yet that is not always the case.

To avoid the high cost of computations involving $\mathbf{A}^{-1}$, one can reduce Σ to $\widehat{\Sigma}$ via ordinary Arnoldi. This approach does match moments of $\mathcal{H}$ and $\widehat{\mathcal{H}}$, but not for the ideal region of the frequency domain. Yet utilizing $\mathcal{K}_k(\mathbf{A}, \mathbf{b})$ avoids the cost of computing with $\mathbf{A}^{-1}$ and in practice often yields $\widehat{\Sigma}$ for which $\widehat{\mathbf{y}} \approx \mathbf{y}$. In particular,

in the current context, one expects that $\widehat{\Sigma}$ defined by (1.8), i.e.,

$$\widehat{\Sigma} := \begin{pmatrix} \mathbf{Q}_k^* \mathbf{A} \mathbf{Q}_k & \mathbf{Q}_k^* \mathbf{b} \\ \mathbf{c} \mathbf{Q}_k & \mathbf{0} \end{pmatrix},$$

yields a reasonable approximation to Σ, where $\mathbf{Q}_k \in \mathbb{C}^{n \times k}$ forms an orthonormal basis for $\mathcal{K}_k(\mathbf{A}, \mathbf{b})$.

A Numerical Experiment

Consider (2.6) with $\beta = 1$ and

$$\begin{aligned} w_0(x) &= 0, \\ f(x,t) &:= \sin(2\pi x)|\sin(2\pi x)|\, e^{-(t-3)} =: f s_1(x) f_2(t) \text{ for } t \geq 0, \\ f(x,t) &:= 0 \text{ for } t < 0. \end{aligned}$$

Recall from the example in Section 2.2.2 that $f_{1,odd} = f_1 \in \mathrm{C}^1(\Omega_2) \setminus \mathrm{C}^2(\Omega_2)$ satisfies Theorem 2.2 (ii) with $m = 3$, where m is defined in the statement of Theorem 2.3 with Theorem 2.1 replaced by Theorem 2.2 (ii).

Re-performing the analysis done in the Theorem 2.3 proof using Theorem 2.2 (ii) rather than Theorem 2.1 reveals that as $j \to \infty$,

$$|s_j(t)| = \mathcal{O}\left(j^{-5}\right), \tag{2.45}$$

where $s_j(t)$ is the jth time-dependent coefficient in the Fourier sine series expansion for the exact solution $w : \Omega \times \mathbb{R} \to \mathbb{R}$ to (2.6), given in (2.7). It follows that

$w : \Omega \times \mathbb{R} \to \mathbb{R}$ is very smooth.

This smoothness is also reflected in the semi-discrete approximation to the continuous solution. In particular, take Σ to be the IVP (2.34)

$$\Sigma := \left(\begin{array}{lcll} \dot{\mathbf{x}} & = & \mathbf{A}\mathbf{x} + \mathbf{b}\mathbf{u} & \text{for } t \geq 0 \\ \mathbf{x} & = & \mathbf{0} & \text{for } t < 0 \\ \mathbf{x}(0) & = & \mathbf{0} & \\ \mathbf{y} & = & \mathbf{c}\mathbf{x} & \text{for all } t \end{array} \right),$$

with $n := 1024$, and

$$\begin{array}{lcl} \mathbf{u}(t) & := & e^{-(t-3)} \text{ for } t \geq 0, \\ \mathbf{u}(t) & := & 0 \text{ for } t < 0, \\ \mathbf{b} & := & \sin(2\pi\mathbf{z})|\sin(2\pi\mathbf{z})|, \\ \mathbf{c} & := & \left(\frac{1}{n+1}\right)^2 \Big(\ 1, \ldots, n \ \Big). \end{array} \tag{2.46}$$

Observe that at non-negative times $\mathbf{y} : \mathbb{R} \to \mathbb{R}$ approximates the trapezoid rule approximation to the continuous quantity $\int_\Omega xw(x,t)dx$ because

$$\begin{array}{rcl} \mathbf{y}(t) & = & \mathbf{c}\mathbf{x}(t) \\ & = & h^2 \sum_{i=1}^{n} i\mathbf{x}(z_i, t) \\ & \approx & \frac{h}{2}\left(z_0 w(0,t) + 2\sum_{i=1}^{n} z_i w(z_i,t) + z_{n+1} w(1,t) \right) \\ & \approx & \int_\Omega xw(x,t)dx. \end{array}$$

This output $\mathbf{y}$ is non-zero roughly over $t \in (0, 35)$ (see Figure 2.8). Note that $\mathbf{c}$ is smooth and satisfies Dirichlet boundary conditions. This system has $|\mathcal{H}(i\omega)(\mathcal{L}\mathbf{u})(i\omega)|$ maximized for ω near 0, as shown in Figure 2.7.

$\widehat{\Sigma}$ obtained from Σ via moment matching by ordinary Arnoldi produces an output $\widehat{\mathbf{y}}(t)$ that approximates $\mathbf{y}(t)$ remarkably well for $k \ll n$ in spite of the fact that it does not match moments of Σ for the optimal region of the frequency domain. See Figure 2.8, which shows that for $k = 150$ and adequately large times within the region $(0, 35)$ (for example $t > 10^{-9}$), the relative error $|\widehat{\mathbf{y}}(t) - \mathbf{y}(t)|/|\mathbf{y}(t)|$ is less than 10^{-3}. If k is decreased to 100, the relative error $|\widehat{\mathbf{y}}(t) - \mathbf{y}(t)|/|\mathbf{y}(t)|$ exceeds 10^{-3} at some times.

From (2.40) and (2.45), for $t \in \{t_i\}_{i=1}^{q}$,

$$\mathbf{x}(t) \;\approx\; \sum_{j=1}^{n} a_j(t)\ \mathbf{v}_j,$$

where

$$|a_j(t)| \;=\; \mathcal{O}\left(j^{-5}\right)$$

as $j \to \infty$ ($j \leq n$, so $j \to \infty$ implies $n \to \infty$), so that the discrete solution $\mathbf{x} : \{t_i\}_{i=1}^{q} \to \mathbb{R}^n$ takes negligible contributions from the high frequency modes of $\mathbf{A}$. Yet $\sigma(\widehat{\mathbf{A}})$ has noticeable influence from high frequency modes for even moderate values of $k \ll 150$ (see Figure 2.9). This high frequency influence in $\sigma(\widehat{\mathbf{A}})$ becomes even more pronounced as k grows. It follows that, although moment matching decreases the model dimension, the stable time-step for the forward Euler scheme is virtually unchanged from that of the full-order model.

Given the smoothness of $\mathbf{x} : \{t_i\}_{i=1}^{q} \to \mathbb{R}^n$ — a sharp $\mathcal{O}(j^{-5})$ decay of the influence by the jth mode — rather than integrate $\widehat{\Sigma}$ to obtain $\widehat{\mathbf{y}}(t) \approx \mathbf{y}(t)$, it is reasonable to consider removing the influence by the high frequency modes in $\widehat{\Sigma}$. One expects that the importance of the eigenmodes of the reduced system $\widehat{\Sigma} \approx \Sigma$ should decline similarly to those of the full-order model Σ, and that, consequently, the highest frequency modes of $\widehat{\Sigma}$ can be ignored via modal reduction without significantly impacting the accuracy of $\widehat{\mathbf{y}}(t) \approx \mathbf{y}(t)$. This expectation proves true experimentally. Section 2.7.2 develops a method for quantifying the importance of each mode $j \in \{1, \ldots, k\}$ belonging to $\widehat{\Sigma}$ and establishes a criterion for choosing a truncation point for modal filtering applied to $\widehat{\Sigma}$ in order to eliminate the high frequency modes from $\widehat{\Sigma}$.

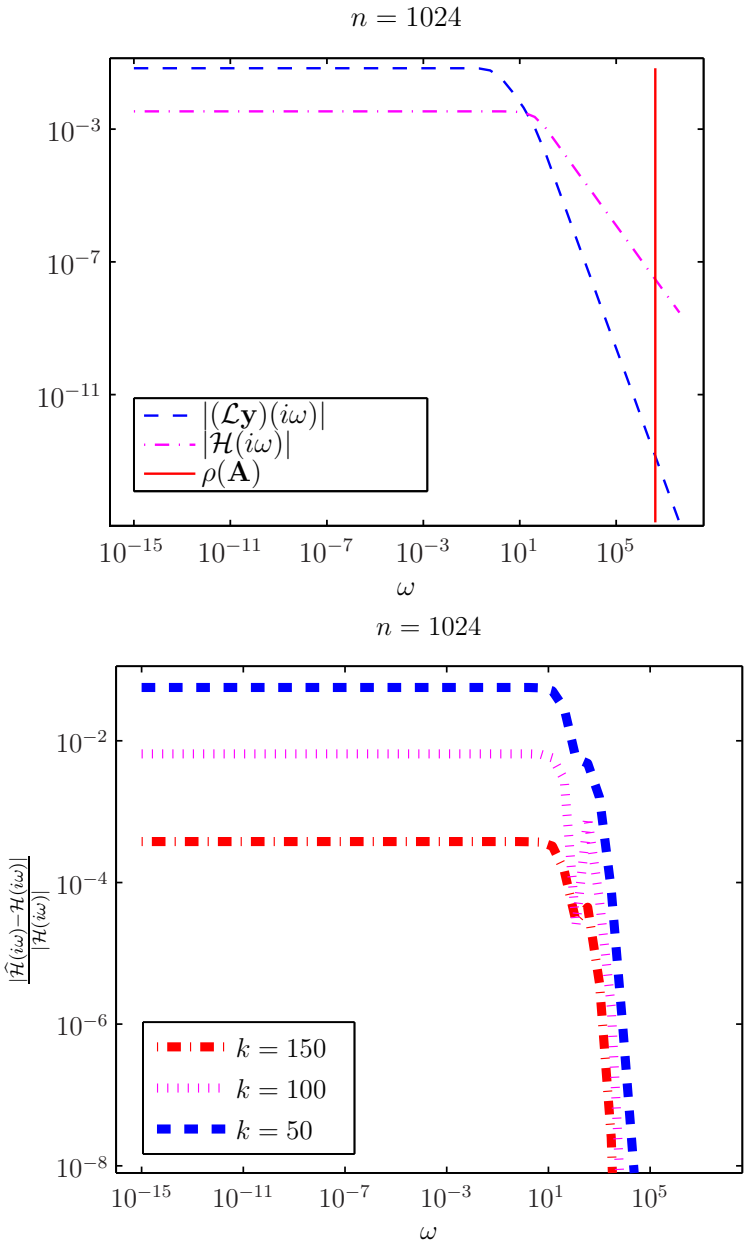


Figure 2.7 : ***Top:*** Superimposed are $|(\mathcal{L}\mathbf{y})(i\omega)| = |\mathcal{H}(i\omega)(\mathcal{L}\mathbf{u})(i\omega)|$ and $|\mathcal{H}(i\omega)|$ versus ω for the full system, Σ, for the example in Sections 2.7.1 and 2.7.2. The system has $\arg\max_{\omega\in\mathbb{R}} |\mathcal{H}(i\omega)(\mathcal{L}\mathbf{u})(i\omega)|$ near 0 with $|\mathcal{H}(i\omega)(\mathcal{L}\mathbf{u})(i\omega)|$ diminishing in size for $|\omega| \gg 0$. Consequently, there is no theoretical expectation that $\widehat{\Sigma}$ obtained via moment matching with ordinary Arnoldi would yield $\widehat{\mathbf{y}}(t) \approx \mathbf{y}(t)$ because it matches Markov parameters. Nonetheless, in practice, $\widehat{\mathbf{y}}(t)$ obtained via ordinary Arnoldi often is a good approximation to $\mathbf{y}(t)$, as is the case in this example. ***Bottom:*** As k increases, the transfer function approximation $\widehat{\mathcal{H}} \approx \mathcal{H}$ becomes more accurate at those frequencies near 0.

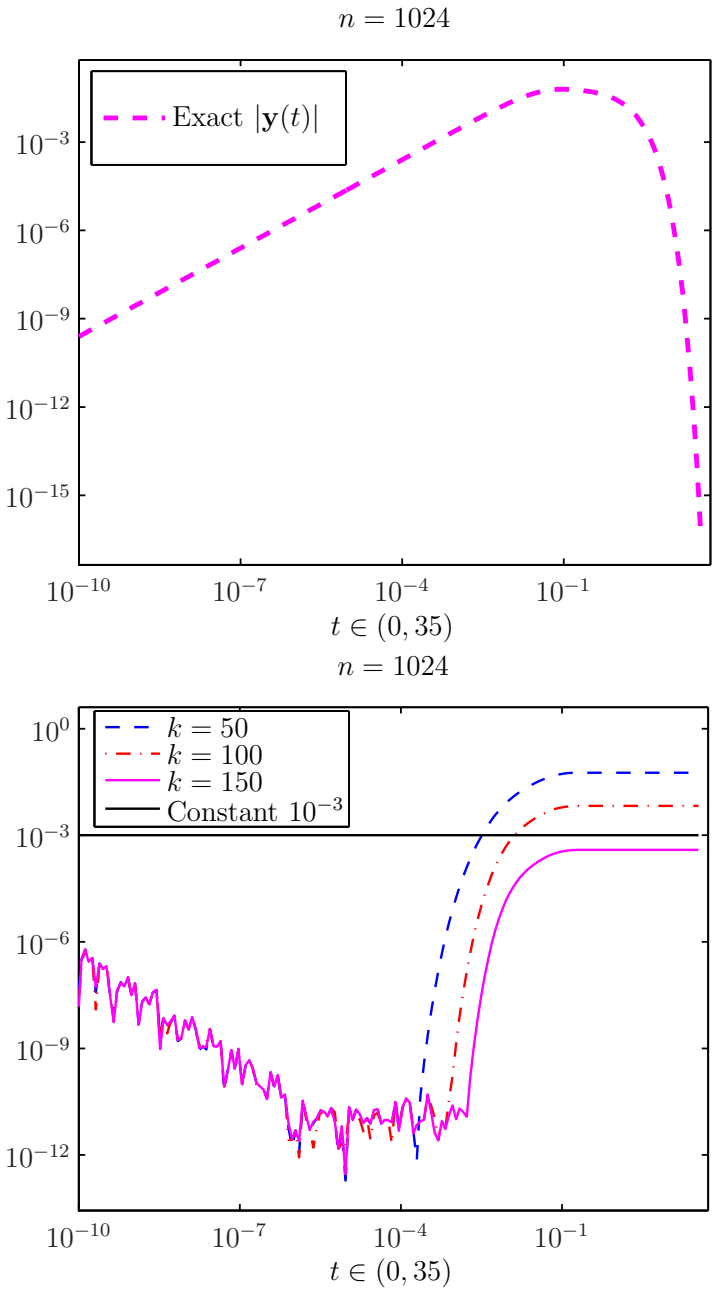


Figure 2.8 : The figures correspond to the example of Section 2.7.1. ***Top:*** Plotted is $|\mathbf{y}(t)|$ (computed exactly) versus t. ***Bottom:*** The figure shows the relative output error $|\widehat{\mathbf{y}}(t) - \mathbf{y}(t)|/|\mathbf{y}(t)|$ versus t for various k. The outputs $\mathbf{y}$ and $\widehat{\mathbf{y}}$ are computed exactly. Notice that the accuracy improves as k increases. Moment matching using ordinary Arnoldi yields accurate output in spite of the fact that it matches the moments of $\mathcal{H}$ and $\widehat{\mathcal{H}}$ at ∞, not the most important region of frequencies in the case of this example. The error curves are jagged because $|\mathbf{y}(t)|$ so is small.

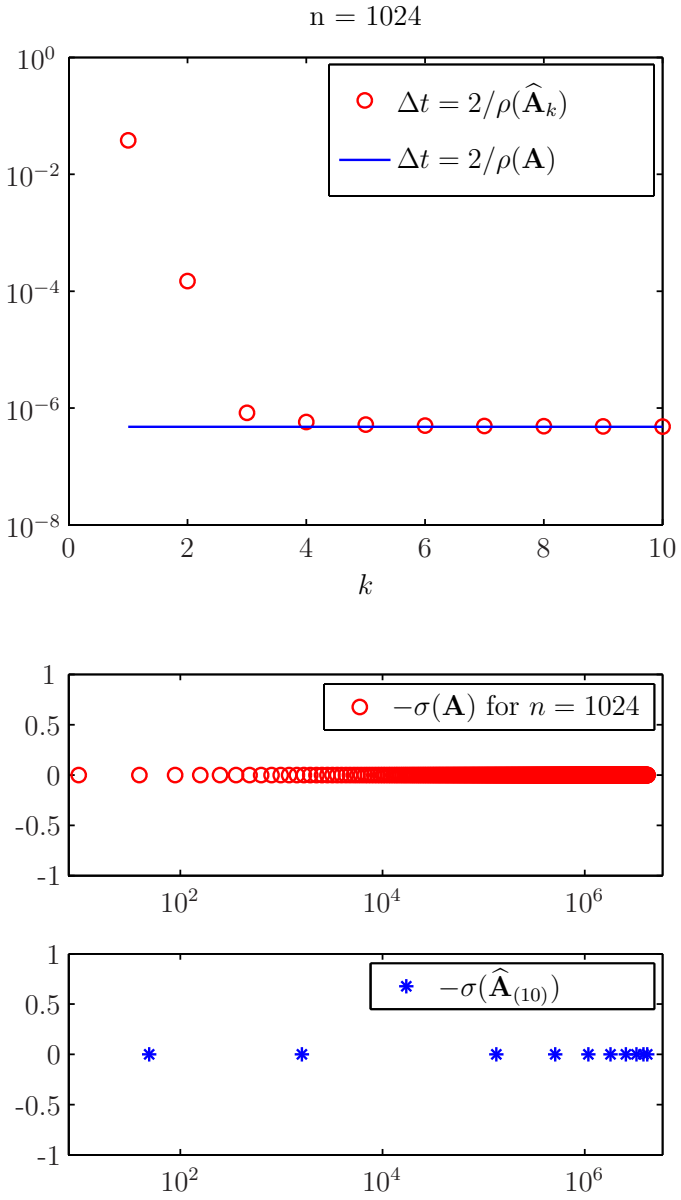


Figure 2.9 : The figures correspond to moment matching via ordinary Arnoldi for the example in Section 2.7.1. ***Top:*** The top diagram shows that $\rho(\widehat{\mathbf{A}}_{(k)})$ is similar in magnitude to $\rho(\mathbf{A})$ by $k = 3$, indicating that, although moment matching has decreased the dimension, the stable time-step for the forward Euler scheme is virtually unchanged. ***Bottom:*** The bottom figure shows $\sigma(\mathbf{A})$ (top) and $\sigma(\widehat{\mathbf{A}}_{(10)})$ (bottom). In particular, note that $\sigma(\widehat{\mathbf{A}}_{(10)})$ is comprised in large part of high frequency Ritz values.

2.7.2 Automated Modal Filtering in Tandem with Moment Matching

Choosing the "Cut-Off" Mode

Upon completing moment matching model reduction to obtain $\widehat{\Sigma}$, diagonalize

$$\widehat{\mathbf{A}} = \widehat{\mathbf{V}}\Theta\widehat{\mathbf{V}}^* \in \mathbb{C}^{k\times k}, \tag{2.47}$$

which is not relatively expensive to compute because $k \ll n$. (Recall that when $\mathbf{A}^* = \mathbf{A}$, Arnoldi simplifies to Lanczos, and $\widehat{\mathbf{A}}^* = \widehat{\mathbf{A}}$ is tridiagonal.) Order the modes from low to high frequency so that

$$\theta_k < \cdots < \theta_1 < 0.$$

Using this decomposition, I derive a natural method for measuring the importance of the $n - l$ highest frequency modes of $\widehat{\Sigma}$.

Observe that $\widehat{\mathbf{b}} \in \mathbb{C}^k$ and $\widehat{\mathbf{c}} \in \mathbb{C}^{1\times k}$ can be written in terms of the eigenvector basis for $\widehat{\mathbf{A}}$. In particular, there exist $\widehat{\mathbf{s}} \in \mathbb{C}^{k\times 1}$ and $\widehat{\mathbf{r}} \in \mathbb{C}^{1\times k}$ such that

$$\begin{aligned}
\widehat{\mathbf{b}} &=: \widehat{\mathbf{V}}\widehat{\mathbf{s}} =: \sum_{j=1}^{k} \widehat{s}_j \widehat{\mathbf{v}}_j \in \mathbb{C}^k; \\
\widehat{\mathbf{c}} &=: \widehat{\mathbf{r}}\widehat{\mathbf{V}}^* =: \sum_{j=1}^{k} \widehat{r}_j \widehat{\mathbf{v}}_j^* \in \mathbb{C}^{1\times k}.
\end{aligned}$$

Observe then that

$$\begin{aligned}
\widehat{\mathcal{H}}(\omega) &:= \widehat{\mathbf{c}}(\omega\mathbf{I} - \widehat{\mathbf{A}})^{-1}\widehat{\mathbf{b}} \\
&= \widehat{\mathbf{r}}\widehat{\mathbf{V}}^*\widehat{\mathbf{V}} \begin{bmatrix} \frac{1}{\omega-\theta_1} & & \\ & \ddots & \\ & & \frac{1}{\omega-\theta_k} \end{bmatrix} \widehat{\mathbf{V}}^*\widehat{\mathbf{V}}\widehat{\mathbf{s}} \\
&= \sum_{j=1}^{k} \frac{\widehat{r}_j\ \widehat{s}_j}{\omega - \theta_j}.
\end{aligned}$$

Consider in particular the application of modal truncation to $\widehat{\Sigma}$ to eliminate all influence of the $k - l$ highest frequency modes, resulting in the system

$$\widetilde{\Sigma} := \begin{pmatrix} \widetilde{\mathbf{A}} & \widetilde{\mathbf{b}} \\ \widetilde{\mathbf{c}} & \mathbf{0} \end{pmatrix},$$

where

$$\begin{aligned}
\widetilde{\mathbf{A}} &:= \begin{pmatrix} \widehat{\mathbf{v}}_1 & \cdots & \widehat{\mathbf{v}}_l \end{pmatrix} \begin{pmatrix} \theta_1 & & \\ & \ddots & \\ & & \theta_l \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{v}}_1^* \\ \vdots \\ \widehat{\mathbf{v}}_l^* \end{pmatrix} \in \mathbb{C}^{k\times k}, \\
\widetilde{\mathbf{b}} &:= \sum_{j=1}^{l} \widehat{s}_j \widehat{\mathbf{v}}_j \in \mathbb{C}^k, \\
\widetilde{\mathbf{c}} &:= \sum_{j=1}^{l} \widehat{r}_j \widehat{\mathbf{v}}_j^* \in \mathbb{C}^{1\times k}.
\end{aligned}$$

Denote the transfer function for $\widetilde{\Sigma}$ by

$$\begin{aligned}\widetilde{\mathcal{H}}(\omega) &:= \widetilde{\mathbf{c}}(\omega\mathbf{I}-\widetilde{\mathbf{A}})^{-1}\widetilde{\mathbf{b}} \\ &= \sum_{j=1}^{l}\frac{\widehat{r}_j\widehat{s}_j}{\omega-\theta_j}.\end{aligned}$$

In particular, for all imaginary $i\omega$,

$$\begin{aligned}\left|\widehat{\mathcal{H}}(i\omega)-\widetilde{\mathcal{H}}(i\omega)\right| &\leq \sum_{j=l+1}^{k}\left|\frac{\widehat{r}_j\widehat{s}_j}{i\omega-\theta_j}\right| \\ &\leq \sum_{j=l+1}^{k}\left|\frac{\widehat{r}_j\widehat{s}_j}{\theta_j}\right| \\ &=: \sum_{j=l+1}^{k}\boldsymbol{\Upsilon}_j,\end{aligned}$$

where entry j of $\boldsymbol{\Upsilon}\in\mathbb{C}^k$ is defined by

$$\boldsymbol{\Upsilon}_j := \frac{|\widehat{r}_j\widehat{s}_j|}{|\theta_j|}. \tag{2.48}$$

Others have derived this bound for $\left|\widehat{\mathcal{H}}(i\omega)-\widetilde{\mathcal{H}}(i\omega)\right|$ in the past (see, e.g., [11, p. 317]).

Observe that a logical modal "cut-off" mode, l, can be chosen by finding $l+1$ such that, for some specified tolerance ϵ,

$$\zeta_{l+1} := \sum_{j=l+1}^{k}\boldsymbol{\Upsilon}_j \leq \epsilon. \tag{2.49}$$

The quantity ζ_{l+1} measures the cumulative importance of modes $\{l+1,\ldots,k\}$ to the

system $\widehat{\Sigma}$.

I will refer to this multi-step dimension reduction algorithm as the *dual-stage reduction algorithm* (step one is moment matching by Arnoldi, followed by modal filtering using (2.49)).

Note that for adequately small k, the vector $\boldsymbol{\Upsilon} \in \mathbb{C}^k$ is inexpensive to compute. In particular, $\widehat{\mathbf{r}} \in \mathbb{C}^{1\times k}$ and $\widehat{\mathbf{s}} \in \mathbb{C}^{k\times 1}$ require only matrix multiplications because

$$\begin{aligned} \mathbf{Q}_k^*\mathbf{b} = \widehat{\mathbf{V}}\widehat{\mathbf{s}} \quad &\Rightarrow \quad \widehat{\mathbf{s}} = \widehat{\mathbf{V}}^*\mathbf{Q}_k^*\mathbf{b}, \\ \mathbf{c}\mathbf{Q}_k = \widehat{\mathbf{r}}\widehat{\mathbf{V}}^* \quad &\Rightarrow \quad \widehat{\mathbf{r}} = \mathbf{c}\mathbf{Q}_k\widehat{\mathbf{V}}. \end{aligned} \tag{2.50}$$

Now in a system in which $\widehat{\mathbf{b}}$ and $\widehat{\mathbf{c}}$ are smooth, i.e., all $|\widehat{s}_j|$ and all $|\widehat{r}_j|$ are small for large j, criterion (2.49) yields a small value for $l+1$, and thus all influences by high frequency modes on $\widehat{\Sigma}$ are removed. Note that, as $\widehat{\mathbf{b}}$ is created from the pair $(\mathbf{A},\ \mathbf{b})$, it is reasonable to expect that when $\mathbf{b}$ is smooth, $\widehat{\mathbf{b}}$ is smooth as well. To reinforce this intuition, recall from (2.43) and (2.50) that

$$\widehat{s}_j = \widehat{\mathbf{v}}_j^*\mathbf{Q}_k^*\mathbf{b} = \langle \mathbf{Q}_k\widehat{\mathbf{v}}_j, \mathbf{V}\mathbf{s} \rangle , \tag{2.51}$$

and realize that $\mathbf{Q}_k\widehat{\mathbf{v}}_j$ is a Ritz vector with corresponding Ritz value θ_j. When the Ritz vector $\mathbf{Q}_k\widehat{\mathbf{v}}_j$ is a good approximation to the eigenvector $\mathbf{v}_m$ for some $m \in \{1,\ldots,n\}$, it follows that

$$\widehat{s}_j \approx s_m. \tag{2.52}$$

The low frequency Ritz vectors generally approximate the low frequency eigenvectors

of Σ to high accuracy. It follows then that $\widehat{s}_j \approx s_j$ for small j. As the lowest frequency modes are the most meaningful to the full-order discrete forcing term, the low frequency modes of the reduced-order model will also be meaningful to the reduced-order discrete forcing term. This argument does not show that the high frequency modes are irrelevant in the reduced-order forcing term whenever the full-order discrete forcing term is smooth. Yet this certainly does hold true in all of my experiments for the semi-discrete heat equation. It follows that, in my experiments, whenever $\mathbf{b}$ is smooth, $\widehat{\mathbf{b}}$ is smooth as well, and thus criterion (2.49) causes the shedding of several modes.

Assume that $\mathbf{u} : \mathbb{R} \to \mathbb{R}$, $\mathbf{u}|_{(-\infty,0)} = 0$, satisfies $(\mathcal{L}\mathbf{u}) \in \mathrm{L}^1(i\mathbb{R})$ with $\|(\mathcal{L}\mathbf{u})\|_{\mathrm{L}^1(i\mathbb{R})} < \mathrm{M} < \infty$. Then by (1.4), if l satisfies (2.49), observe that $\widetilde{\mathbf{y}} \approx \widehat{\mathbf{y}}$ has $\mathcal{O}(\epsilon)$ point-wise accuracy as $\epsilon \to 0$ because

$$\begin{aligned} \|\widehat{\mathbf{y}} - \widetilde{\mathbf{y}}\|_{\mathrm{L}^\infty(\mathbb{R})} &= \max_{t \geq 0} \left\{ \frac{1}{2\pi} \left| \int_{\mathbb{R}} \left[\widehat{\mathcal{H}}_k(i\omega) - \widetilde{\mathcal{H}}(i\omega) \right] (\mathcal{L}\mathbf{u})(i\omega)\, e^{i\omega t}\, d\omega \right| \right\} \\ &\leq \frac{\epsilon \mathrm{M}}{2\pi}. \end{aligned}$$

Implementation of Modal Truncation

Notice that an implementation based upon the preceding derivation does not reduce the dimension during the modal filtration step as did the implementation of (1.13) in Section 1.3. Nonetheless, the two implementations yield mathematically equivalent state variables and outputs. The preceding implementation is useful for the derivation

of an intelligent choice of l according to $\zeta_{l+1} \leq \epsilon$. Yet it should not be implemented in the current setting of $\widehat{\mathbf{A}} = \widehat{\mathbf{A}}^*$, which has a well-conditioned eigenvector basis $\widehat{\mathbf{V}}$, because one can instead attain a system with dimension $l < k$ without concerns due to an ill-conditioned eigenvector basis.

Specifically, in the notation of Section 2.7.2, the system Σ_3 in the implementation of Section 1.3 has the form

$$\widetilde{\Sigma} := \left(\begin{array}{lcll} \dot{\boldsymbol{\chi}}_1 & = & \Theta_1 \boldsymbol{\chi}_1 + \widehat{\mathbf{V}}_1^* \widehat{\mathbf{b}} \mathbf{u} & \text{for } t \geq 0 \\ \boldsymbol{\chi}_1 & = & \mathbf{0} & \text{for } t < 0 \\ \boldsymbol{\chi}_1(0) & = & \widehat{\mathbf{V}}_1^* \widehat{\mathbf{x}}(0) & \end{array} \right) \tag{2.53}$$

where $\boldsymbol{\chi}_1 : \mathbb{R} \to \mathbb{R}^l$,

$$\Theta_1 := \left(\begin{array}{ccc} \theta_1 & & \\ & \ddots & \\ & & \theta_l \end{array} \right) \in \mathbb{R}^{l \times l},$$

and

$$\widehat{\mathbf{V}}_1 := \left(\begin{array}{ccc} \widehat{\mathbf{v}}_1 & \ldots & \widehat{\mathbf{v}}_l \end{array} \right) \in \mathbb{R}^{k \times l}.$$

Recall that, upon integrating (2.53) for $\boldsymbol{\chi}_1$, one attains $\widetilde{\mathbf{x}} : \mathbb{R} \to \mathbb{R}^k$ such that $\widetilde{\mathbf{x}} \approx \widehat{\mathbf{x}}$ by defining

$$\widetilde{\mathbf{x}} := \widehat{\mathbf{V}}_1 \boldsymbol{\chi}_1$$

and $\widetilde{\mathbf{y}} : \mathbb{R} \to \mathbb{R}$ such that $\widetilde{\mathbf{y}} \approx \widehat{\mathbf{y}}$ by defining

$$\widetilde{\mathbf{y}} := \widehat{\mathbf{c}}\widetilde{\mathbf{x}}$$

(see (1.14)).

This definition of $\widetilde{\Sigma}$ (2.53) not only filters out the high frequency eigenmodes but also has dimension $l < k$, smaller than that of $\widehat{\Sigma}$. Thus this is the definition that should be utilized in implementations.

Return to the Numerical Experiment of Section 2.7.1

To illustrate this technique for modal truncation in practice, I return to the numerical experiment of Section 2.7.1. Suppose that the application at hand requires the final

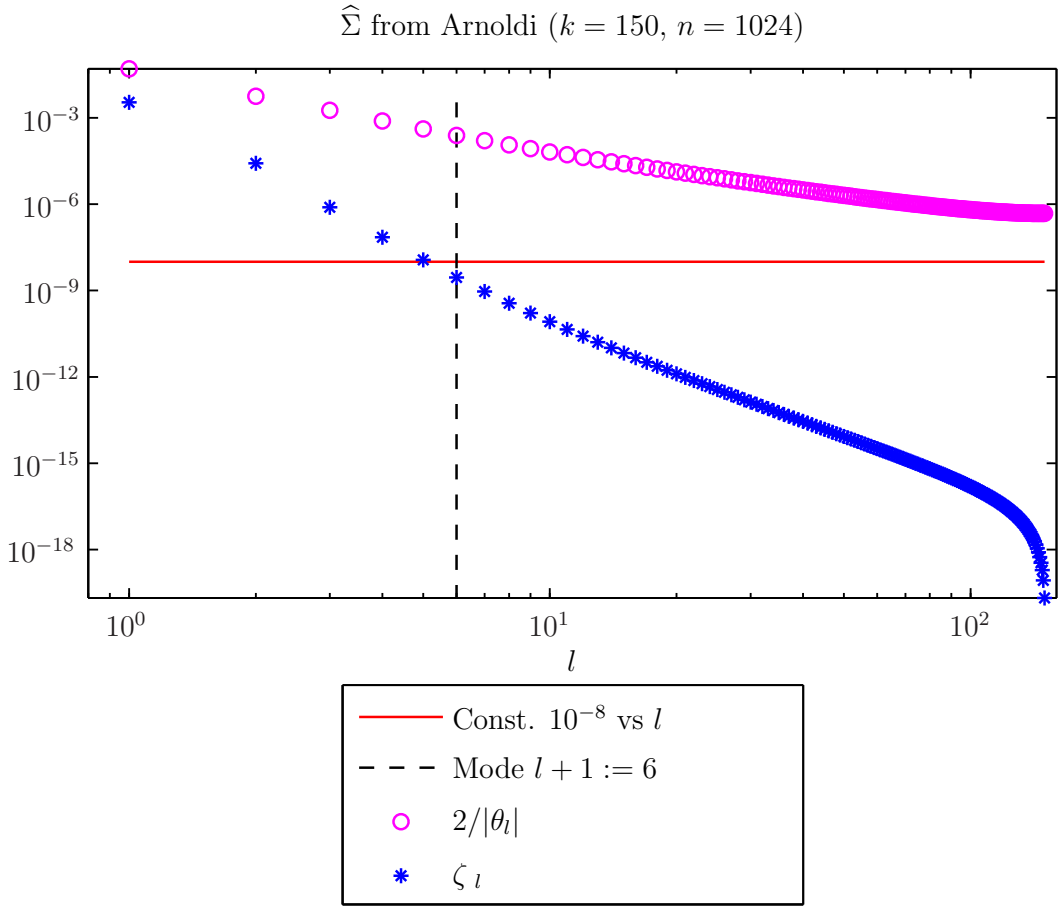


Figure 2.10 : The blue stars and magenta circles show ζ_l (2.49) and $2/|\theta_l|$ (stable time-step for integration of $\dot{\widetilde{\mathbf{x}}} = \widetilde{\mathbf{A}}\widetilde{\mathbf{x}}$) versus l, respectively, where $\widehat{\Sigma}$ was obtained by moment matching with Arnoldi for the numerical experiment in Sections 2.7.1 and 2.7.2. The solid red line shows the cutoff threshold, $\epsilon := 10^{-8}$. The dashed black line marks mode $l+1 = 6$, the mode of lowest index j such that $\zeta_j \leq \epsilon$. Thus $l = 5$ is the truncation mode.

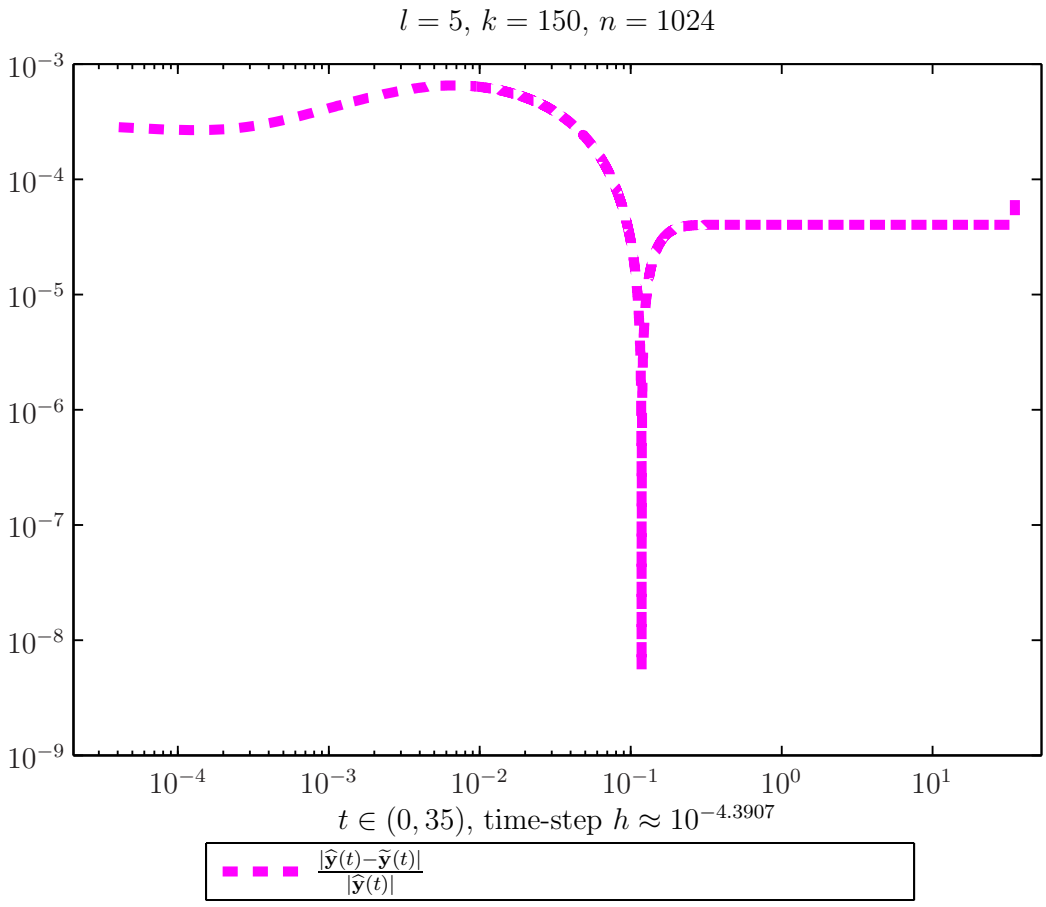


Figure 2.11 : For the example in Sections 2.7.1 and 2.7.2, the plot shows the left hand side of (2.56) versus t, indicating that $\widetilde{\Sigma}$ of dimension only $l := 5$ satisfies (2.56) with $\delta := 10^{-3}$. $\widehat{\mathbf{y}}(t)$ was computed exactly, while $\widetilde{\mathbf{y}}(t)$ was approximated by using the forward Euler scheme with step size $h := 1/(5\rho(\widetilde{\mathbf{A}})) \approx 10^{-4.3907}$. The accuracy of $\widetilde{\mathbf{y}}(t)$ is remarkable considering that it was generated with a time-step that is roughly two orders of magnitude larger than the stable time-step for integration of $\dot{\widehat{\mathbf{x}}} = \widehat{\mathbf{A}}\widehat{\mathbf{x}}$ by the forward Euler scheme and that the dimension is reduced by two orders of magnitude from Σ to $\widetilde{\Sigma}$.

approximation to satisfy

$$\frac{|\mathbf{y}(t) - \widetilde{\mathbf{y}}(t)|}{|\mathbf{y}(t)|} \leq c\delta \tag{2.54}$$

for some specified δ and some constant c on the order of 1 for all times t such that $\mathbf{y}(t) \neq 0$. Then by selecting k such that

$$\frac{|\widehat{\mathbf{y}}(t) - \mathbf{y}(t)|}{|\mathbf{y}(t)|} < \delta, \tag{2.55}$$

(2.54) follows by requiring that the relative error between the outputs of $\widehat{\Sigma}$ and $\widetilde{\Sigma}$ be small, i.e.,

$$\frac{|\widehat{\mathbf{y}}(t) - \widetilde{\mathbf{y}}(t)|}{|\widehat{\mathbf{y}}(t)|} < \delta. \tag{2.56}$$

That (2.55) and (2.56) are sufficient for (2.54) follows from the observation that

$$|\widehat{\mathbf{y}}(t)| \leq |\mathbf{y}(t)| + \delta|\mathbf{y}(t)|.$$

Consider in particular $\delta := 10^{-3}$. As was pointed out in Figure 2.8, $k = 150$ satisfies (2.55) with $\delta = 10^{-3}$ at all adequately large times (before the output decays to negligible levels), namely $t > 10^{-9}$. It follows that, as the time-step for the forward Euler scheme h is much greater than 10^{-9}, if the forward Euler approximation to $\widetilde{\mathbf{y}}(t)$ satisfies (2.56) at all times in the discrete grid, then it will also satisfy (2.54) at all times in the discrete grid.

As computation of $\widehat{\mathbf{y}}(t)$ is much faster than computation of $\mathbf{y}(t)$, I will determine whether (2.54) is satisfied by computing the left hand side of (2.56) and measuring whether (2.56) is satisfied, which I can do by the preceding reasoning. I select $\epsilon := 10^{-8}$ as the cut-off tolerance, which gives $l = 5$, as highlighted by Figure 2.10, which

shows ζ_l versus l. The resulting $\widetilde{\Sigma}$ then satisfies (2.56) (see Figure 2.11), so that $\widetilde{\mathbf{y}}(t)$ attains the required order of 10^{-3} relative accuracy in (2.54).

Not only does $\widetilde{\Sigma}$ satisfy the accuracy requirements, but it does so using a time-step that is two orders of magnitude larger than the stable time-step for integration of $\dot{\widehat{\mathbf{x}}} = \widehat{\mathbf{A}}\widehat{\mathbf{x}}$ by the forward Euler scheme and with the dimension reduced by two orders of magnitude from Σ to $\widetilde{\Sigma}$. This improvement in time-step required for 10^{-3} relative accuracy (2.56) renders integration by the forward Euler scheme palatable. (Note also the improvement in the stable forward Euler step-size by roughly three orders of magnitude — from $2/\rho(\widehat{\mathbf{A}}) \approx 10^{-6.3225}$ for $\widehat{\Sigma}$ to $2/\rho(\widetilde{\mathbf{A}}) \approx 10^{-3.3907}$ for $\widetilde{\Sigma}$.) In addition to providing improvements in time-step, the modal truncation step allows one to further reduce the model to grasp the essence of the system (the dimension 5 system $\widetilde{\Sigma}$ encapsulates the dimension 1024 system Σ). Modal filtering in tandem with moment matching works remarkably well for this example.

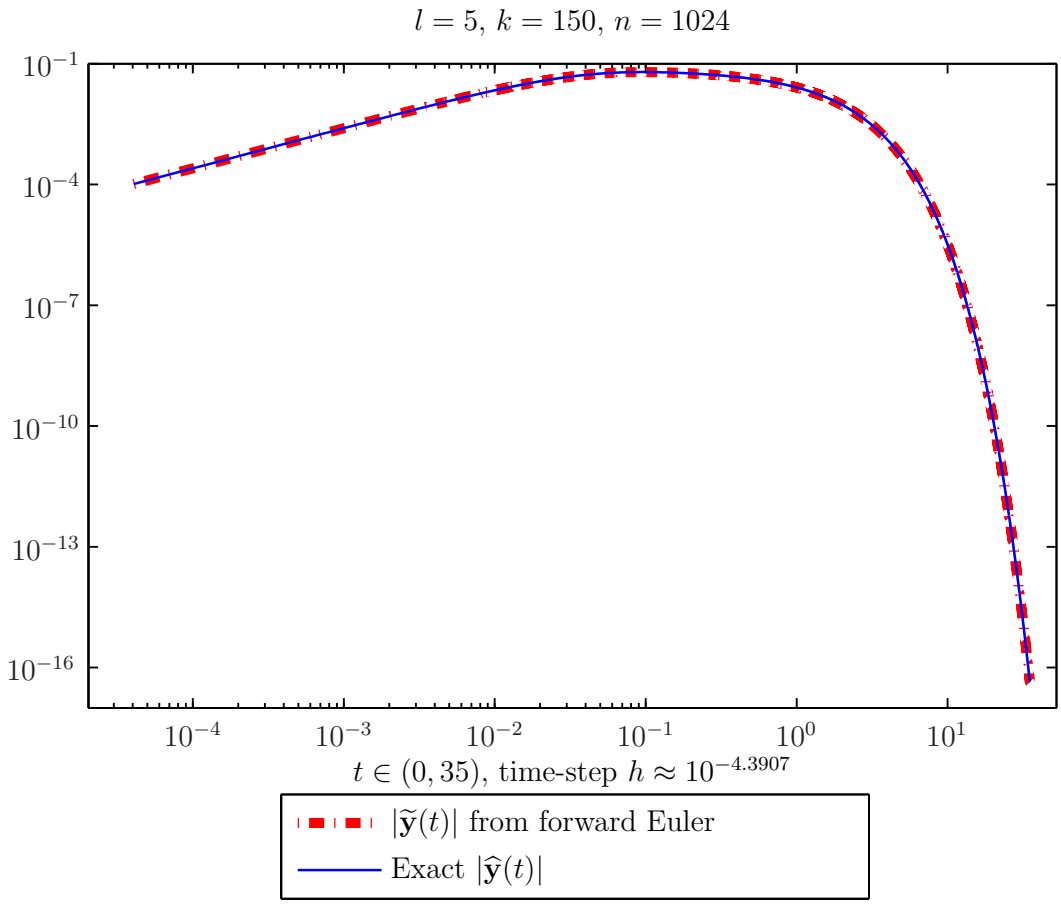


Figure 2.12 : For the example in Sections 2.7.1 and 2.7.2, the plot superimposes $|\widetilde{\mathbf{y}}(t)|$ and $|\widehat{\mathbf{y}}(t)|$. $\widehat{\mathbf{y}}(t)$ was computed exactly, while $\widetilde{\mathbf{y}}(t)$ was approximated by using the forward Euler scheme with step size $h := 1/(5\rho(\widetilde{\mathbf{A}})) \approx 10^{-4.3907}$.

2.8 Conclusions

The smoothness of odd extensions of the forcing term $f : \Omega \times \mathbb{R} \to \mathbb{R}$ at all times $t \geq 0$ directly correlates to the relevance of high frequency eigenmodes to the exact solution $w : \Omega \times \mathbb{R} \to \mathbb{R}$ to the heat equation if the initial state is zero. Of note are estimates on the Fourier sine series coefficients $\{s_j(t)\}_{j\in\mathbb{N}}$ for the series expansion

$$w(x,t) = \sum_{j\in\mathbb{N}} s_j(t)\phi_j(x). \tag{2.57}$$

Such decay estimates reveal that, when $f\big|_{\partial\Omega} = 0$, the smoother functions in the set $\{\big[f(\cdot,r)\big]_{odd} : \Omega_2 \to \mathbb{R}\}_{r\in[0,T]}$ are, the more rapidly the coefficients $\{s_j(t)\}_j$ decay with j, and the less relevant the higher frequency modes are to the exact solution.

This theory for the continuous system generalizes immediately to the discrete realm. In particular, the jth discrete mode of the discrete system approximates the jth continuous mode of the continuous system in the sense that the discrete eigenvalue approximates the continuous eigenvalue, and the piecewise linear interpolant defined by the jth discrete eigenvector evaluated on the finite difference grid $\mathbf{z} \in \mathbb{R}^n$ converges in L^2 to a multiple of the jth continuous eigenfunction. Given the connection between the jth discrete and continuous modes, it is intuitive that the decay estimates for continuous modes can be used to estimate the decay of contributions by the corresponding discrete modes. In particular, if the state variable $\mathbf{x}(t_i) \approx u(\mathbf{z}, t_i)$ approximates the continuous solution on the finite difference spatial grid at a collection

of discrete times $t_i \in \{t_m\}_{m=1}^{q}$, then for adequately large n,

$$\mathbf{x}(t_i) \approx \sum_{j=1}^{n} a_j(t_i)\mathbf{v}_j.$$

Under these circumstances, any estimate for the decay of $|s_j(t_i)|$ (2.57) in j yields a corresponding estimate for the decay of $|a_j(t_i)|$ in j by the relationship $a_j(t) = \sqrt{\frac{2}{\beta}} c_j s_j(t)$. Most notably, as $f_{odd} : \Omega_2 \times \mathbb{R} \to \mathbb{R}$ becomes smoother, so does the discrete approximation $\mathbf{x} : \mathbb{R} \to \mathbb{R}^n$ to $w : \Omega \times \mathbb{R} \to \mathbb{R}$. Such discrete decay estimates reveal that the discrete modes that determine the strict upper bound on the stable time-step for the forward Euler scheme are also virtually irrelevant. Their omission results in negligible loss of accuracy in the state variable, while improving the stable time-step by orders of magnitude.

Theoretically, one could apply modal filtering directly to the full-sized system to both reduce dimension and improve the stable time-step by orders of magnitude. Yet for large systems, the cost of diagonalization required for modal filtering renders it unjustifiable. Nonetheless, the cost of modal truncation applied to a reduced-order model generated via moment matching is more acceptable in the context of the heat equation. In the numerical experiment considered here, the output generated by moment matching closely matches the output originating from the full-order model for reduced dimension much smaller than the full-order dimension.

When the linear, time-invariant, heat system Σ is to be integrated using an explicit scheme, moment matching acting alone leaves much to be desired. Indeed, while

moment matching reduces the dimension, it does not generally reduce the stable forward Euler time-step. A dual-stage reduction that combines moment matching with modal filtering is far more effective. The modal phase is intelligently informed by the relevance of each eigenmode to the reduced-order model. In particular, a logical cut-off mode l is selected by requiring that the reduced-order transfer function and its modally filtered counterpart differ by at most some quantity $\epsilon > 0$ everywhere on the imaginary axis. This can be accomplished by requiring that

$$\zeta_l \leq \epsilon \tag{2.58}$$

because the parameter ζ_l, given in (2.48) and (2.49), satisfies

$$\max_{\omega \in \mathbb{R}} \left| \widehat{\mathcal{H}}(i\omega) - \widetilde{\mathcal{H}}(i\omega) \right| \leq \zeta_l.$$

One can calculate the whole set of parameter values $\{\zeta_l\}_{l=1}^{k}$ at low computational cost in the context of the heat equation because $\mathbf{A} = \mathbf{A}^*$.

In the setting of a linear, time-invariant system originating from a heat IBVP, if the full-order discrete forcing term is smooth, then the reduced-order discrete forcing term obtained by moment matching is typically smooth as well. Given sufficient smoothness in the reduced-order forcing term, modal filtration determined according to (2.58) eliminates the influence of high frequency eigenmodes, alleviating the strict upper bound on the stable time-step for integration of $\dot{\widehat{\mathbf{x}}} = \widehat{\mathbf{A}}\widehat{\mathbf{x}}$ by the forward Euler

scheme. Moreover, the modal filtration step identifies and retains only the most relevant modes of $\widehat{\Sigma}$. In effect, the dual-stage procedure builds a system $\widetilde{\Sigma}$ that encapsulates the system Σ using a much smaller dimension. Bearing (2.58) in mind with

$$\mathbf{y}(t) - \widetilde{\mathbf{y}}(t) \;=\; \frac{1}{2\pi}\int_{\mathbb{R}} (\mathcal{L}\mathbf{u})(i\omega) e^{i\omega t}\left(\mathcal{H}(i\omega) - \widetilde{\mathcal{H}}(i\omega)\right) d\omega,$$

the resulting approximation $\widetilde{\mathbf{y}} \approx \mathbf{y}$ is highly accurate.

Chapter 3

The Convection-Diffusion Equation

Chapter 2 shows that the dual-stage reduction algorithm functions well for the heat equation. Yet that setting enjoys the benefit of symmetry and well-conditioned eigenvectors. Highly nonnormal differential operators present major challenges to the performance of moment matching in tandem with modal filtering. Chapter 3 presents preliminary experiments demonstrating the reduced effectiveness of the dual-stage procedure under such circumstances. Before investigating two numerical examples, Sections 3.1 and 3.2 introduce the continuous and semi-discrete forms of the convection-diffusion equation, respectively.

3.1 Solution to the Continuous Problem

Recall the definition of $\mathcal{Q}$ from (2.1), and define $\mathrm{H} : \mathcal{Q} \to \mathrm{L}^2(\Omega)$ by

$$\mathrm{H}\, g := d \bigtriangleup g + c\, g_x, \tag{3.1}$$

which I refer to as the *convection-diffusion operator* with convection and diffusion coefficients c and d, respectively. For fixed $T > 0$, the Dirichlet *convection-diffusion equation* is the IBVP

$$\begin{array}{lcl} w_t & = & \mathrm{H}w + f \quad \text{on} \quad \Omega \times [0,T], \\ w|_{\Omega\times\{0\}} & = & w_0, \\ w|_{\partial\Omega\times[0,T]} & = & 0, \\ w|_{\Omega\times(-\infty,T)} & = & 0 \end{array} \tag{3.2}$$

for given $f : \Omega \times (-\infty, 0) \to \mathbb{R}$, $f|_{\Omega\times(-\infty,0)} = 0$, and $w_0 : \Omega \to \mathbb{R}$.

Observe that H is not self-adjoint in $\langle \cdot, \cdot \rangle_\Omega$, but an equivalent IBVP with self-adjoint $\widehat{\mathrm{H}} : \mathcal{Q} \to \mathrm{L}^2(\Omega)$ is quickly derived [15, p. 16-17, 28-29]. Namely,

$$\begin{array}{lcl} \widehat{w}_t & = & \widehat{\mathrm{H}}\ \widehat{w} + \widehat{f} \quad on \quad \Omega \times [0,T], \\ \widehat{w}|_{\Omega\times\{0\}} & = & \widehat{w}_0, \\ \widehat{w}|_{\partial\Omega\times[0,T]} & = & 0, \\ \widehat{w}|_{\Omega\times(-\infty,0)} & = & 0, \end{array} \tag{3.3}$$

where

$$\begin{array}{rcl} \widehat{w} & := & e^{cx/(2d)} w, \\ \widehat{w}_0 & := & e^{cx/(2d)} w_0, \\ \widehat{f} & := & e^{cx/(2d)} f, \end{array}$$

and for any $g \in \mathcal{Q}$,

$$\widehat{\mathrm{H}}\ g := d \ \triangle\ g - \frac{c^2}{4d}\ g. \tag{3.4}$$

$\widehat{\mathrm{H}}$ has eigenvalues given by

$$\widehat{\lambda}_j \ := \ -d \left(\tfrac{\pi j}{\beta}\right)^2 - \tfrac{c^2}{4d}, \quad j \in \mathbb{N}\ . \tag{3.5}$$

As the eigenfunctions of $\widehat{\mathrm{H}}$ (identical to those of L from (2.2)),

$$\phi_j(x) \;:=\; \sqrt{\tfrac{2}{\beta}}\sin\left(\tfrac{\pi j x}{\beta}\right), \quad j \in \mathbb{N}, \tag{3.6}$$

form an orthonormal basis for $\mathcal{Q}$, one can solve for $\widehat{w}$ in (3.3) using a Fourier series expansion. One readily observes that

$$\widehat{w}(x,t) = \sum_{j\in\mathbb{N}} a_j(t)\phi_j(x), \tag{3.7}$$

with

$$\begin{aligned} a_j(t) \;&:=\; \langle \widehat{w}(\cdot,t), \phi_j\rangle_\Omega \\ &=\; e^{\widehat{\lambda}_j t}a_j(0) + \textstyle\int_0^t e^{(t-s)\widehat{\lambda}_j} b_j(s)ds, \end{aligned} \tag{3.8}$$

where $b_j(s) := \left\langle \widehat{f}(\cdot,s), \phi_j \right\rangle_\Omega$ and $a_j(0) := \langle \widehat{w}(\cdot,0), \phi_i\rangle_\Omega$.

Throughout Chapter 3 I will assume that f satisfies $\lim_{t\to\infty}\sup_{x\in\Omega}|\widehat{f}(x,t)| = 0$ and $\|f\|_{\mathrm{L}^\infty(\Omega\times\mathbb{R})} < \infty$, together sufficient for the conclusion that for all j,

$$\lim_{t\to\infty}\left\langle \widehat{f}(x,t), \phi_j\right\rangle_\Omega \;=\; \left\langle \lim_{t\to\infty}\widehat{f}(x,t), \phi_j\right\rangle_\Omega = 0$$

(see, e.g., [16, Proposition 6, p. 84]). In particular, it follows that contributions by the symmetrized forcing term $\widehat{f}$ to solution $\widehat{w}$ decay to 0 as $t\to\infty$. Moreover, I assume throughout Chapter 3 that $d\pi^2/\beta > -c/(4d)$, which is sufficient to insure that $\max_j \widehat{\lambda}_j < 0$.

*Theorem 3.1 (**Eigenvalues and Eigenfunctions of** H)*

The convection-diffusion operator $\mathrm{H} : \mathcal{Q} \to \mathrm{L}^2(\Omega)$ *of* (3.1) *has eigenfunctions*

$$\phi_j(x) := e^{-\frac{cx}{2d}} \sin\left(\frac{j\pi x}{\beta}\right), \qquad j \in \mathbb{N}, \tag{3.9}$$

with corresponding eigenvalues given by

$$\lambda_j := -\left(\frac{(2d\pi j/\beta)^2 + c^2}{4d}\right). \tag{3.10}$$

Moreover, $\{\phi_j\}_{j\in\mathbb{N}}$ *is a set of linearly independent functions on* Ω*.*

For a proof, see Appendix A.3.

Remark: Notice that $\sigma(\mathrm{H}) \in (-\infty, 0)$ assuming that $d > 0$ and $c \in \mathbb{R}$. I will assume this to be the case throughout Chapter 3.

3.2 Semi-Discretization Using Centered Finite Differences

Section 3.2 introduces the semi-discrete convection-diffusion equation, yielding the nonnormal discrete convection-diffusion operator, $\mathbf{A}$.

Take $\mathbf{A}_D \in \mathbb{R}^{n\times n}$ to be the discrete *diffusion* operator, i.e., the discrete Laplacian from Section 2.4, with corresponding finite difference grid $\mathbf{z} \in \mathbb{R}^n$ (2.21) and $h := \frac{\beta}{n+1}$. The discrete *convection* operator is analogously defined via the centered finite

difference approximation for the first derivative. Namely, for all $m \in \{1, \ldots, n\}$,

$$w'(z_m) \;=\; \frac{1}{2h}\left(w(z_{m+1}) - w(z_{m-1})\right) + \mathcal{O}(h^2)$$

as $n \to \infty$ (see, e.g., [12, p. 4]). Denote

$$\mathbf{A}_C := \frac{1}{2h}\begin{pmatrix} 0 & 1 & & & \\ -1 & & & & \\ & & \ddots & & \\ & & & & 1 \\ & & & -1 & 0 \end{pmatrix} \in \mathbb{R}^{n\times n}.$$

If $w \in \mathcal{Q}$, then $\|\mathbf{A}_C w(\mathbf{z}) - w'(\mathbf{z})\| = \mathcal{O}(h^{3/2})$ as $n \to \infty$.

The discrete approximation to the convection-diffusion operator (3.1) is $\mathbf{A} \in \mathbb{R}^{n\times n}$ given by

$$\begin{aligned} \mathbf{A} \;&:=\; d\mathbf{A}_D + c\mathbf{A}_C \\ &= \begin{pmatrix} -\frac{2d}{h^2} & \left(\frac{d}{h^2} + \frac{c}{2h}\right) & & & \\ \left(\frac{d}{h^2} - \frac{c}{2h}\right) & & & & \\ & & \ddots & & \\ & & & & \left(\frac{d}{h^2} + \frac{c}{2h}\right) \\ & & & \left(\frac{d}{h^2} - \frac{c}{2h}\right) & -\frac{2d}{h^2} \end{pmatrix} \\ &=: \begin{pmatrix} a_0(n) & a_1(n) & & & \\ a_{-1}(n) & & & & \\ & & \ddots & & \\ & & & & a_1(n) \\ & & & a_{-1}(n) & a_0(n) \end{pmatrix}. \end{aligned} \tag{3.11}$$

In general, $\mathbf{A}^*\mathbf{A} \neq \mathbf{A}\mathbf{A}^*$, and thus, unlike the discrete Laplacian, $\mathbf{A}$ is not unitarily

diagonalizable. One can show that if $a_{-1}(n) \neq 0 \neq a_1(n)$, then

$$\sigma(\mathbf{A}) = \left\{ \frac{-2d}{h^2} + \frac{2}{h}\sqrt{\frac{d^2}{h^2} - \frac{c^2}{4}} \cos\left(\frac{j\pi}{n+1}\right) \right\}_{j=1}^{n} \tag{3.12}$$

are the eigenvalues of $\mathbf{A}$ [12, p. 277]. Fix c and d and select n adequately large such that $\sigma(\mathbf{A}) \subset \mathbb{R}$. Then index the eigenvalues of $\mathbf{A}$ so that

$$\mu_n < \cdots < \mu_1 < 0,$$

where $\mu_1 < 0$ follows from the assumptions that $c \in \mathbb{R}$ and $d > 0$. The corresponding eigenvectors of $\mathbf{A}$ are given by

$$\mathbf{v}_j := \frac{1}{c_j} \begin{bmatrix} \left(\frac{a_{-1}(n)}{a_1(n)}\right)^{1/2} \sin\left(\frac{j\pi}{n+1}\right) \\ \vdots \\ \left(\frac{a_{-1}(n)}{a_1(n)}\right)^{n/2} \sin\left(\frac{nj\pi}{n+1}\right) \end{bmatrix}, \tag{3.13}$$

where c_j is chosen so that $\|\mathbf{v}_j\| = 1$ [12, p. 277]. This results in the (non-unitary) diagonalization

$$\mathbf{A} = \mathbf{V}\mathbf{M}\mathbf{V}^{-1}. \tag{3.14}$$

As in (2.20), assume $f : \Omega \times \mathbb{R} \to \mathbb{R}$ satisfies $f(x,t) = f_1(x)f_2(t)$ for some $f_1 : \Omega \to \mathbb{R}$ and $f_2 : \mathbb{R} \to \mathbb{R}$, $f_2|_{(-\infty,0)} = 0$. In this context, the semi-discretized version of the IBVP (3.2) is

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{A}\mathbf{x} + \mathbf{b}\mathbf{u} \quad \text{for } t \geq 0, \\ \mathbf{x} &= \mathbf{0} \quad \text{for } t < 0, \\ \mathbf{x}(0) &= w_0(\mathbf{z}), \end{aligned} \tag{3.15}$$

where $\mathbf{A}$ and $\mathbf{b} := f_1(\mathbf{z})$ are constant, and $\mathbf{x}_m(t) \approx w(z_m, t)$ for all $m \in \{1, \ldots, n\}$.

3.3 Dual-Stage Dimension Reduction of the Semi-Discretized Problem

I now generalize the dual-stage algorithm to non-symmetric systems Σ and show its impracticality for problems for which moment matching by non-inverted Arnoldi is ineffective.

Non-Hermitian Implementation

In the scenario that

$$\mathbf{A} = \mathbf{V}\mathbf{M}\mathbf{V}^{-1}$$

is non-symmetric with $\sigma(\mathbf{A}) \subset \mathbb{R}$, moment matching via Arnoldi (Section 2.7.1) and the subsequent modal filtering phase (Section 2.7.2) of the dual-stage procedure are performed in an identical fashion to the scenario in which $\mathbf{A} \in \mathbb{R}^{n \times n}$ is Hermitian, with the replacement of $\mathbf{V}^* \in \mathbb{C}^{n \times n}$ and $\widehat{\mathbf{V}}^* \in \mathbb{C}^{k \times k}$ by $\mathbf{V}^{-1} \in \mathbb{C}^{n \times n}$ and $\widehat{\mathbf{V}}^{-1} \in \mathbb{C}^{k \times k}$, respectively. Given these changes, one goes about defining

$$\zeta_{l+1} := \sum_{j=l+1}^{k} (\boldsymbol{\Upsilon})_j \in \mathbb{R} \tag{3.16}$$

and the related modal filtration criterion, $\zeta_{l+1} \leq \epsilon$, in identical fashion to that of the Hermitian case.

Computationally, the dual-stage procedure is slightly more expensive in the non-Hermitian case, yet still palatable when k is adequately small. In particular, $\widehat{\mathbf{V}}$ and Θ from (2.47) can be computed using `eig`, but $\widehat{\mathbf{s}} = \widehat{\mathbf{V}}^{-1}\widehat{\mathbf{b}}$ (rather than $\widehat{\mathbf{V}}^{*}\widehat{\mathbf{b}}$) must be computed as well. Nonetheless, it can be efficiently approximated by `Matlab`'s `backslash` command via $\widehat{\mathbf{V}}^{-1}\widehat{\mathbf{b}} = \widehat{\mathbf{V}}\backslash\widehat{\mathbf{b}}$. Vector $\widehat{\mathbf{r}} \in \mathbb{C}^{1\times k}$ is no costlier to compute than in the Hermitian case — requiring two matrix-vector multiplications — as shown in (2.50).

3.3.1 A Numerical Experiment

When moment matching by ordinary Arnoldi does not adequately reduce the system dimension to subsequently perform modal filtering, the dual-stage procedure cannot be justified in terms of cost. While moment matching by *shift-inverted* Arnoldi may be more effective in reducing the dimension in such scenarios, one would not generally use integration by the forward Euler scheme for such problems, and thus a modal phase becomes less relevant if one's main purpose in using the dual-stage procedure is to improve the stable time-step. Specifically, if one can justify inverting $(\mathbf{A}-\sigma\mathbf{I}) \in \mathbb{R}^{n\times n}$ to create $\widehat{\mathbf{A}} \in \mathbb{R}^{k\times k}$ using shift-inverted Arnoldi, then usually one can also justify inverting $(\mathbf{I} - h\widehat{\mathbf{A}}) \in \mathbb{R}^{k\times k}$ to integrate $\widehat{\Sigma}$ by the *backward Euler scheme*,

$$\begin{aligned} \widehat{\mathbf{x}}_0 &= \widehat{\mathbf{x}}(0), \\ \widehat{\mathbf{x}}_{m+1} &:= (\mathbf{I} - h\widehat{\mathbf{A}})^{-1}(\widehat{\mathbf{x}}_m + h\widehat{\mathbf{b}}\mathbf{u}(t_{m+1})). \end{aligned} \tag{3.17}$$

For this scheme, all time-steps $h > 0$ are stable if $\sigma(\widehat{\mathbf{A}}) \subset (-\infty, 0)$. Note, however, that if one seeks the maximal possible dimension reduction of Σ, then modal filtering following a moment matching step via shift-inverted Arnoldi can certainly add value.

To illustrate an example for which moment matching by ordinary Arnoldi is ineffective in reducing the dimension, consider the IVP

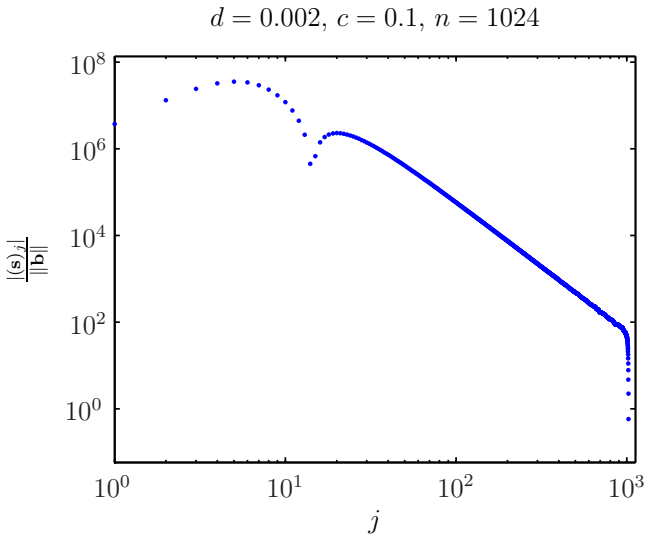


Figure 3.1 : Plotted is $|(\mathbf{s})_j|/\|\mathbf{b}\|$ versus j for the example in Section 3.3.1. Notice that these coefficients are large for both small and high frequency modes, a reflection of the nonnormality of $\mathbf{A}$. One expects that for $\widehat{\Sigma} \approx \Sigma$, the coefficients $|(\widehat{\mathbf{s}})_j| = |(\widehat{\mathbf{V}}^{-1}\widehat{\mathbf{b}})_j|$ would behave similarly.

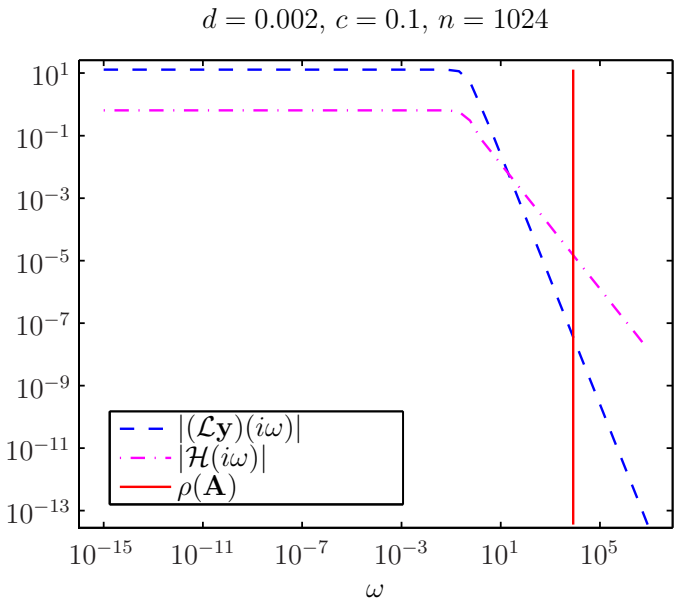


Figure 3.2 : Superimposed are $|(\mathcal{L}\mathbf{y})(i\omega)| = |\mathcal{H}(i\omega)(\mathcal{L}\mathbf{u})(i\omega)|$ and $|\mathcal{H}(i\omega)|$ versus ω for the full system, Σ, for the example in Section 3.3.1. Notice that the system has $|\mathcal{H}(i\omega)(\mathcal{L}\mathbf{u})(i\omega)|$ maximized for ω near 0. Consequently, as is the case for the heat equation example in Chapter 2, there is no theoretical expectation that $\widehat{\Sigma}$ obtained via ordinary Arnoldi would yield $\widehat{\mathbf{y}}(t) \approx \mathbf{y}(t)$. Yet for the semi-discrete heat equation, moment matching through ordinary Arnoldi does obtain $\widehat{\mathbf{y}}(t) \approx \mathbf{y}(t)$ for $k \ll n$. Such favorable behavior does not hold for this highly nonnormal convection-diffusion equation, as shown in Figure 3.3.

$$\left(\begin{array}{lcll} \dot{\mathbf{x}} & = & \mathbf{A}\mathbf{x} + \mathbf{b}\mathbf{u} & \text{for } t \geq 0 \\ \mathbf{x} & = & \mathbf{0} & \text{for } t < 0 \\ \mathbf{y} & = & \mathbf{c}\mathbf{x} & \\ \mathbf{x}(0) & = & \mathbf{0} & \end{array}\right) \tag{3.18}$$

with $\mathbf{A}$ given by (3.11) and diffusion and convection coefficients $d := 0.002$ and $c := 0.1$, respectively. Take

$$\begin{aligned}
\mathbf{b} &:= \sin(2\pi\mathbf{z})|\sin(2\pi\mathbf{z})| &&=: \mathbf{V}\mathbf{s}, \\
\mathbf{c} &:= \left(\tfrac{1}{n+1}\right)^2 \left(1, \ldots, n \right) &&=: \mathbf{r}\mathbf{V}^{-1}, \\
\mathbf{u}(t) &:= e^{3-t} \text{ for } t \geq 0, \\
\mathbf{u}(t) &:= 0 \text{ for } t < 0,
\end{aligned} \tag{3.19}$$

identical to $\mathbf{b}$, $\mathbf{c}$ and $\mathbf{u}$ in the example in Sections 2.7.1 and 2.7.2.

For this choice of convection and diffusion constants, $\mathbf{A}$ is significantly nonnormal, or, equivalently, the basis given by $\mathbf{V}$ is far from orthogonal [19, p. 115-118]. Consequently, contributions by most of the higher frequency modes to $\mathbf{b}$ relative to the size of $\mathbf{b}$ tend to be large, as shown in Figure 3.1.

Recall from Section 2.7.1 that for any Σ such that $|(\mathcal{L}\mathbf{y})(i\omega)|$ is largest for frequencies ω near 0 — as is the case for this IVP (see Figure 3.2) — optimal moment matching requires the moments near 0 of $\widehat{\Sigma}$ and Σ to match, i.e., (2.44). This is achieved by matching moments through inverted Arnoldi, which is costly given that it utilizes $\mathbf{A}^{-1}$.

For the example semi-discretized heat equation in Chapter 2, moment matching through ordinary Arnoldi attains $\widehat{\mathbf{y}}(t) \approx \mathbf{y}(t)$ for $k \ll n$ in spite of the fact that the Laplace transform of its output has maximum magnitude for frequencies $i\omega$ near 0. Thus, in that setting, modal filtering can be applied to $\widehat{\Sigma}$ at a much lower cost than that required to apply it directly to Σ.

Yet for this convection-diffusion equation, moment matching by ordinary Arnoldi generates $\widehat{\mathbf{y}}(t)$ that is not nearly as accurate for modest k. Figure 3.3 shows that in

order to achieve

$$\frac{|\widehat{\mathbf{y}}(t) - \mathbf{y}(t)|}{|\mathbf{y}(t)|} < 10^{-3} \tag{3.20}$$

for $n = 1024$ at all times such that $\mathbf{y}(t) \neq 0$, one must select k to be of the same order of magnitude as n. (Times $t \in (0, 40)$ are plotted because $\mathbf{y} : \mathbb{R} \to \mathbb{R}$ decays in magnitude below 10^{-12} at times beyond 40.) The requirement that k be so large relative to n in order to satisfy (3.20) is unsurprising in light of the poor approximation given by the ϵ-pseudospectra of $\widehat{\mathbf{A}}$ to those of $\mathbf{A}$ for k such that $k \ll n$. (The ϵ-pseudospectrum of a matrix $\mathbf{D} \in \mathbb{R}^{n \times n}$, $\sigma_\epsilon(\mathbf{D})$, is defined by [19]

$$\sigma_\epsilon(\mathbf{D}) := \{z \in \mathbb{C} : z \in \sigma\left(\mathbf{D} + \mathbf{E}\right) \text{ for some } \mathbf{E} \text{ with } \|\mathbf{E}\| \leq \epsilon\}.) \tag{3.21}$$

Figure 3.4 shows that $\{\sigma_\epsilon(\widehat{\mathbf{A}})\}_{\epsilon \in \{10^0, 10^{-2}, \ldots, 10^{-10}\}}$ do not resemble their counterparts in the set $\{\sigma_\epsilon(\mathbf{A})\}_{\epsilon \in \{10^0, 10^{-2}, \ldots, 10^{-10}\}}$ until k is on the order of n. That is, k must be on the order of n before $\widehat{\mathbf{A}}$ begins to accurately reflect the nonnormality of $\mathbf{A}$.

Yet as k approaches n, modal filtering of $\widehat{\Sigma}$ is almost as costly as modal filtering of Σ. The dual-stage reduction scheme would fail to be cost-effective in this example for large n.

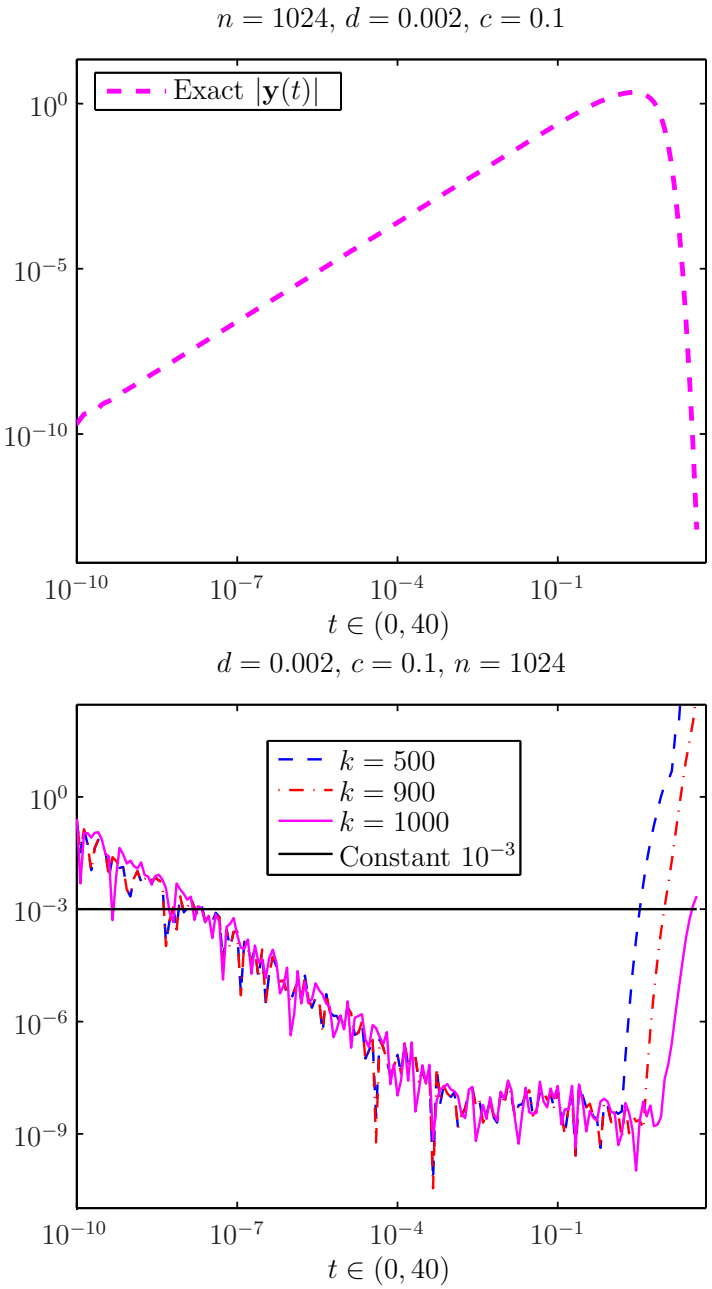


Figure 3.3 : Both figures correspond to the example of Section 3.3.1. ***Top:*** Plotted is $|\mathbf{y}(t)|$ (computed exactly) versus t. ***Bottom:*** The figure shows relative output error $|\widehat{\mathbf{y}}(t) - \mathbf{y}(t)|/|\mathbf{y}(t)|$ versus t for various k. Outputs $\mathbf{y}(t)$ and $\widehat{\mathbf{y}}(t)$ are computed exactly. Not until $k \approx 10^3$ does $\widehat{\mathbf{y}}$ satisfy (3.20). Yet modal filtering applied to $\widehat{\Sigma}$ with a dimension on the same order of magnitude as that of Σ is not generally justifiable. Error curves are jagged because $|\mathbf{y}(t)|$ is small.

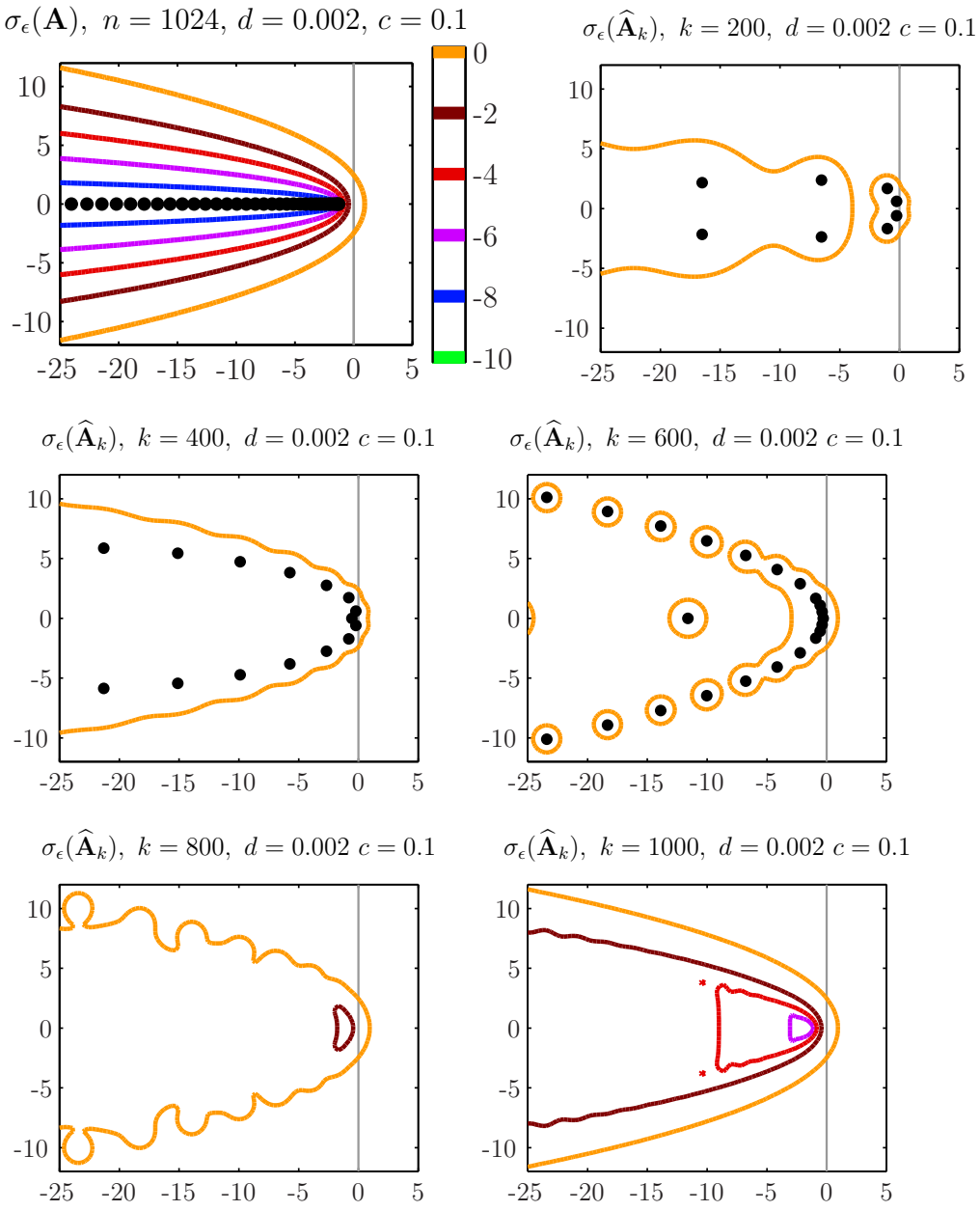


Figure 3.4 : For the example in Section 3.3.1, shown are the pseudospectra of $\mathbf{A} \in \mathbb{R}^{n\times n}$ (*top left*, $n := 1024$) and $\widehat{\mathbf{A}} \in \mathbb{R}^{k\times k}$ for $k \in \{200,\ 400,\ 600,\ 800,\ 1000\}$. All five plots use an identical color-coding scheme to that given in the top left plot. Black dots are the eigenvalues — omitted in the plots for $k = 800$ and 1000, where they overshadow $\sigma_{10^{-2}}(\widehat{\mathbf{A}})$, $\sigma_{10^{-4}}(\widehat{\mathbf{A}})$ and $\sigma_{10^{-6}}(\widehat{\mathbf{A}})$. Notice that the pseudospectra of $\widehat{\mathbf{A}}$ that are shown here do not closely resemble their counterparts for $\mathbf{A}$ until $k \approx 1000$. In the plots shown here, $\sigma_{10^{-4}}(\widehat{\mathbf{A}})$ and $\sigma_{10^{-6}}(\widehat{\mathbf{A}})$ are not even visible for $k < 1000$.

3.3.2 A Convection-Diffusion Example for which Moment Matching *is* Feasible

While the dual-stage reduction algorithm is not practical for large n in the example of Section 3.3.1, that is not the case for all convection-diffusion systems. In the example of Section 3.3.1, the dual-stage algorithm is rendered ineffective by the pseudospectra grazing the imaginary axis for small $\epsilon > 0$. To illustrate this point, consider the identical IVP with differential operator shifted by $\alpha > 0$. That is,

$$\left(\begin{array}{lcll} \dot{\mathbf{x}} & = & (\mathbf{A} - \alpha\mathbf{I})\,\mathbf{x} + \mathbf{bu} & \text{for } t \geq 0 \\ \mathbf{x} & = & \mathbf{0} & \text{for } t < 0 \\ \mathbf{y} & = & \mathbf{cx} & \\ \mathbf{x}(0) & = & \mathbf{0} & \end{array}\right), \tag{3.22}$$

where $\mathbf{A}$, $\mathbf{b}$, $\mathbf{c}$, $\mathbf{u}$, $d := .002$, $c := .1$ and $n := 1024$ are identical to those of the example in Section 3.3.1. Notice that the only difference between this and the example of Section 3.3.1 is that α shifts $\sigma(\mathbf{A})$ along the real axis. Hence I refer to (3.22) as the *shifted convection-diffusion equation.*

When the shift is $\alpha = 5$, moment matching through ordinary Arnoldi produces $\widehat{\mathbf{y}}$ that satisfies (3.20) for $k = 150 \ll n$, as shown in Figure 3.5. (Times $t \in (0, 35)$ are shown because the output magnitude decays to levels below 10^{-15} at times beyond 35.) I select $k := 150$. In this setting, modal filtering following the moment matching step is cost effective and can significantly improve the stable time-step for integration of the problem $\dot{\widehat{\mathbf{x}}} = \mathbf{A}\widehat{\mathbf{x}}$ by the forward Euler scheme. Yet stability does not imply

accuracy, and the time-step required to yield an output $\widetilde{\mathbf{y}} : \mathbb{R} \to \mathbb{R}$, $\widetilde{\mathbf{y}}|_{(-\infty,0)} = 0$, that satisfies

$$\frac{|\widehat{\mathbf{y}}(t) - \widetilde{\mathbf{y}}(t)|}{|\widehat{\mathbf{y}}(t)|} < 10^{-3} \tag{3.23}$$

for times such that $\widehat{\mathbf{y}}(t) \neq 0$, is only moderately improved from $\widehat{\Sigma}$ to $\widetilde{\Sigma}$.

For example, choosing tolerance $\epsilon := 10^{-4.5}$ and requiring that l satisfy

$$\zeta_{l+1} \leq \epsilon$$

yields $l = 6 \ll k \ll n$ (see Figure 3.6). The resulting system $\widetilde{\Sigma}$ has stable time-step given by $2/[\rho(\widetilde{\mathbf{A}})] \approx 10^{-0.9535}$, almost three orders of magnitude larger than that of $\widehat{\Sigma}$, $2/[\rho(\widehat{\mathbf{A}})] \approx 10^{-3.6237}$. Yet the time-step required for attaining order 10^{-3} relative accuracy (3.23) is roughly $10^{-3.15}$ (see Figure 3.7). That is an improvement by roughly half of an order of magnitude over a time-step $10^{-3.624} < 10^{-3.6237} \approx 2/\rho(\widehat{\mathbf{A}})$ that attains the same level of accuracy when using forward Euler to integrate $\widehat{\Sigma}$ without modal filtering. Hence the improvement in the time-step required for attaining the desired level of accuracy is not nearly as significant as the improvement observed in the stable time-step.

Moreover, to reap the maximal, albeit moderate, improvement in the time-step that produces 10^{-3} relative accuracy, one has to select just the right tolerance ϵ to attain the optimal value of l, as other values of l yield diminished improvements in the accurate time-step. I demonstrate this trend in Figure 3.8. The delicate

nature of selecting l in order to attain the optimal improvement in the time-step that produces the desired accuracy level is emphasized by the observation that, upon filtering only two additional modes beyond $l = 6$, i.e., to $l = 4$, the time-step required for 10^{-3} relative accuracy plummets unfavorably — even below that required when using forward Euler to integrate $\widehat{\Sigma}$ without filtering. Hence, for this example, there is at best a moderate improvement in the time-step required to attain (3.23).

Nonetheless, the dual-stage procedure still carries value in this setting. Specifically, if one does not eliminate too many modes through the second reduction step, then the dual-stage procedure generates a system $\widetilde{\Sigma}$ with a dimension on the order of 10 that encapsulates the dimension 1024 system Σ. That dimension reduction may alone justify the dual-stage reduction.

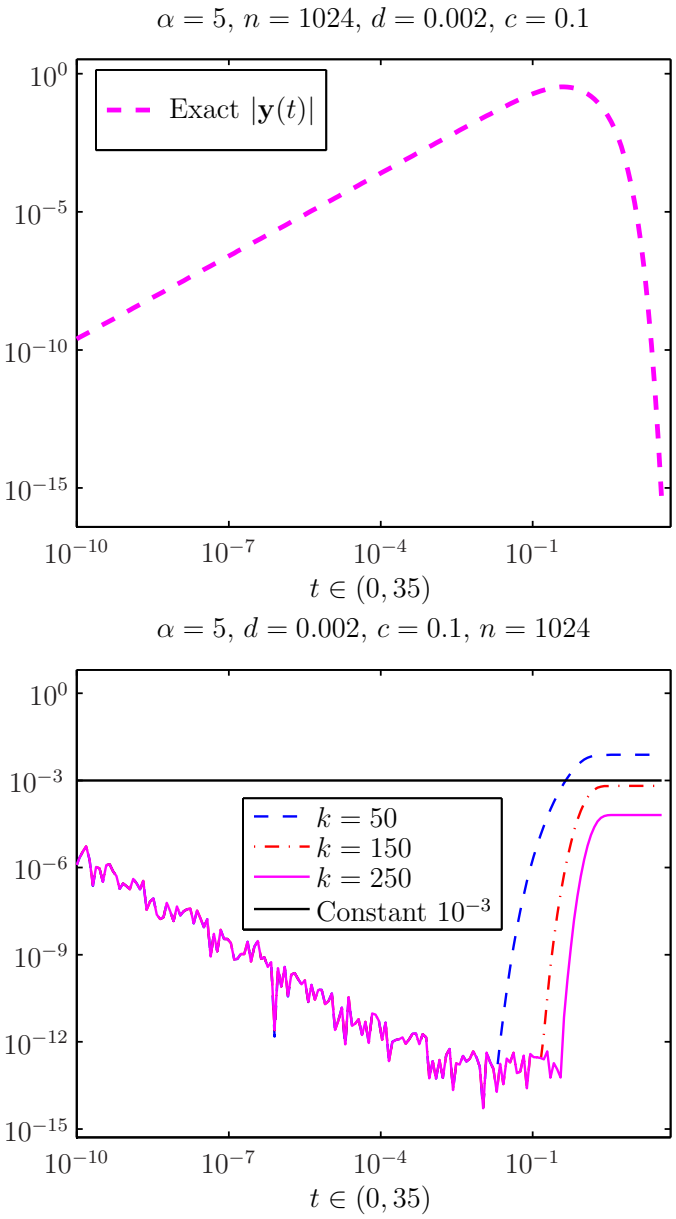


Figure 3.5 : These plots correspond to the example of Section 3.3.2. ***Top:*** Plotted is $|\mathbf{y}(t)|$ (computed exactly) versus t. ***Bottom:*** The figure shows $|\widehat{\mathbf{y}}(t) - \mathbf{y}(t)|/|\mathbf{y}(t)|$ versus t for various k with $\mathbf{y}(t)$ and $\widehat{\mathbf{y}}(t)$ computed exactly. For $k = 150$, one order of magnitude smaller than n, $\widehat{\mathbf{y}}$ satisfies (3.20) at adequately large times. As moment matching reduces the dimension so significantly, modal filtering can be applied to $\widehat{\Sigma}$ in a relatively efficient manner while maintaining the desired relative accuracy on the order of 10^{-3}, assuming that the 10^{-3} level of relative accuracy is maintained in $\widetilde{\mathbf{y}} \approx \widehat{\mathbf{y}}$ produced via forward Euler. The error curves are jagged because $|\mathbf{y}(t)|$ is small.

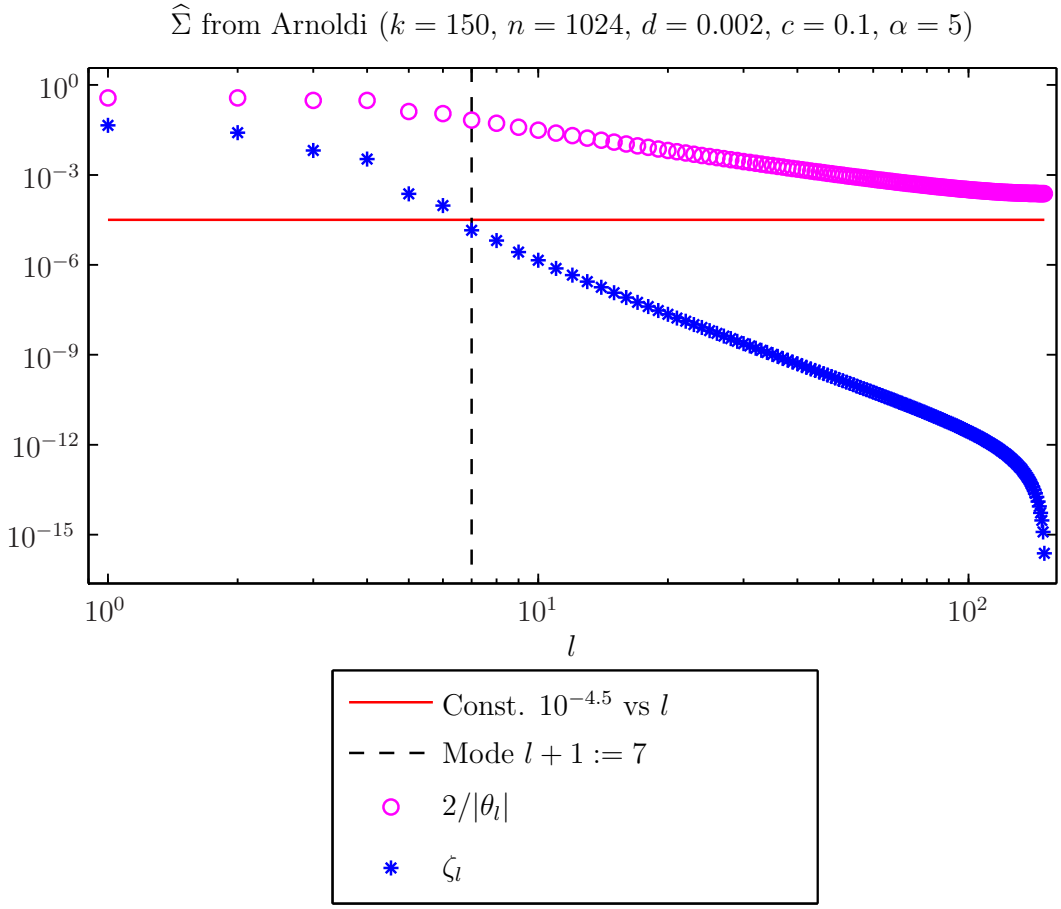


Figure 3.6 : For the example in Section 3.3.2, blue stars show ζ_l versus l. Magenta circles show $2/|\theta_l|$ versus l, the stable time-step for the forward Euler scheme. Evidently, the collective importance of modes $\{j, \ldots, k\}$ in the system $\widehat{\Sigma}$ (measured by ζ_j) decays significantly with j.

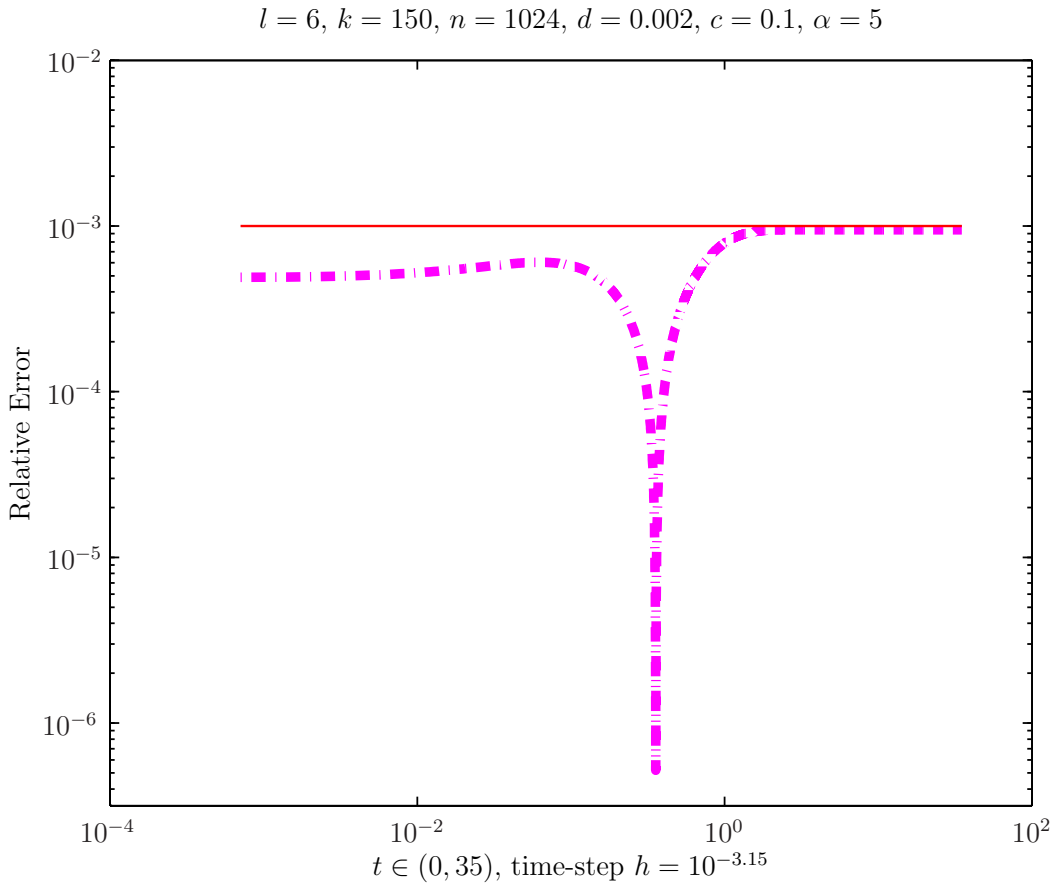


Figure 3.7 : For the example in Section 3.3.2, plotted is relative error $|\widehat{\mathbf{y}}(t) - \widetilde{\mathbf{y}}(t)|/|\widehat{\mathbf{y}}(t)|$ versus t, indicating that $\widetilde{\Sigma}$ of dimension only $l := 6$ satisfies (3.23). $\widehat{\mathbf{y}}(t)$ was computed exactly, while $\widetilde{\mathbf{y}}(t)$ was approximated by using the forward Euler scheme with the step size $h := 10^{-3.15}$.

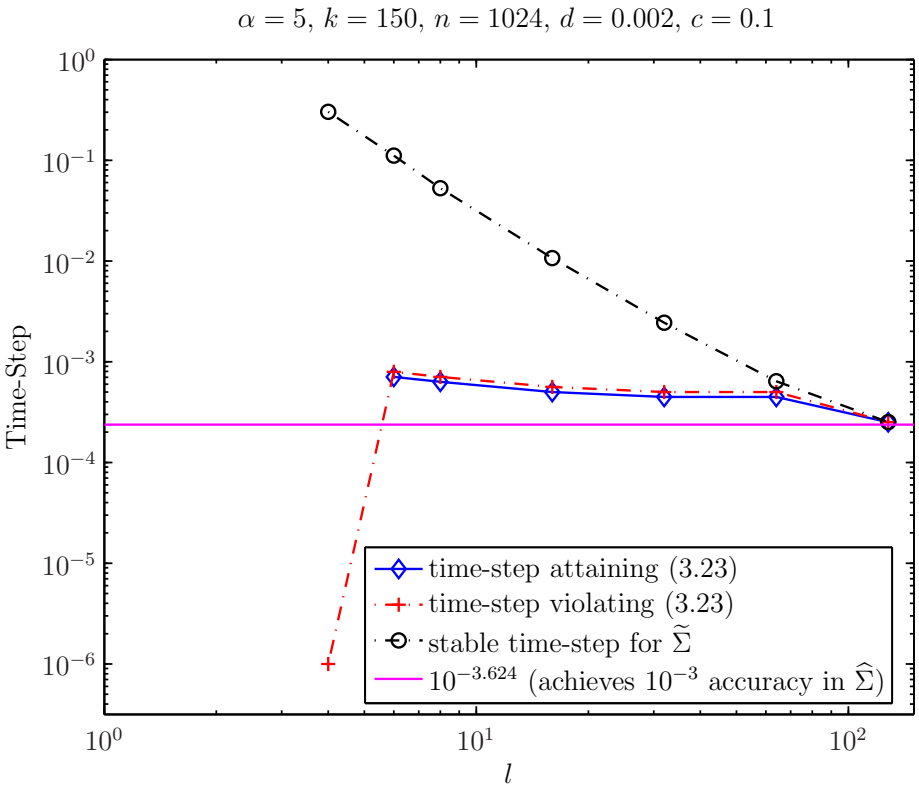


Figure 3.8 : For the example in Section 3.3.2, for several values of the truncation mode l, three time-steps are shown:

- One attaining (3.23) in $\widetilde{\Sigma}$ (blue diamonds, not shown for $l = 4$);
- one violating (3.23) in $\widetilde{\Sigma}$ (red crosses);
- and the stability limit for the system $\widetilde{\Sigma}$ (black circles).

Observe that for $l \ll k$, the stable time-step for $\widetilde{\Sigma}$ is orders of magnitude larger than that required for attaining (3.23). Selecting $l = 6$ yields $\widetilde{\Sigma}$ for which forward Euler with the time-step $h = 10^{-3.15}$ produces $\widetilde{\mathbf{y}}$ satisfying (3.23). That is an improvement by almost half of an order of magnitude over a time-step $10^{-3.624} < 10^{-3.6237} \approx 2/\rho(\widehat{\mathbf{A}})$ with which forward Euler attains the same level of accuracy when integrating $\widehat{\Sigma}$ without filtering.
Upon filtering only two modes beyond $l = 6$, i.e., to $l = 4$, the time-step required to satisfy (3.23) for $\widetilde{\Sigma}$ plummets unfavorably beyond that required to produce 10^{-3} relative accuracy when applying forward Euler to $\widehat{\Sigma}$ without filtering.

3.4 Conclusions

While the dual-stage algorithm readily generalizes to systems for which $\mathbf{A}$ is non-Hermitian, it has reduced effectiveness in the numerical experiments given here. For the example in Section 3.3.1, in order to capture the full nonnormality of the full-order system, the reduced system formed by moment matching must have a dimension on the same order as the dimension of the full-order model. Given this requirement, modal reduction is not advisable because its computational cost is not significantly improved by the first reduction step.

Shifting the spectrum of $\mathbf{A}$ by -5 away from the imaginary axis improves the performance of moment matching by ordinary Arnoldi in reducing the full-size IVP, as shown by the example in Section 3.3.2. Yet the subsequent modal truncation step has mixed results in this setting. While modal truncation substantially improves the *stable* time-step, it only moderately improves the time-step required to attain relative accuracy of 10^{-3}. Yet the modal filtration stage is highly effective in reducing the dimension of $\widehat{\Sigma}$. While the improvement in the time-step that produces 10^{-3} relative accuracy does not alone justify the modal phase in this setting, the substantial added dimension reduction due to the modal phase can do so.

Thus, in the less ideal setting of the highly nonnormal convection-diffusion operator, the dual stage algorithm has a mixed performance. Such settings can demand alternative reduction techniques.

Chapter 4

Concluding Remarks

The two objectives of the dual-stage reduction algorithm are dimension reduction and alleviation of the stringent upper bound on the stable time-step for integration of linear, time-invariant systems by explicit numerical integration schemes. This algorithm works particularly well in the setting of the heat equation but deteriorates in setting of the the convection-diffusion problem.

I conclude by examining the major points of Chapters 2 and 3 in a broader context, and I discuss open questions related to this research.

The Setting of the Heat Equation

In the setting of the heat equation, classical estimates for complex Fourier series coefficients lead to estimates for Fourier sine series coefficients of smooth functions. Most notably, when the forcing term $f : \Omega \times \mathbb{R} \to \mathbb{R}$, $f|_{\Omega\times(-\infty,0)} = 0$, satisfies $f(x,t) = f_1(x)f_2(t)$, where $f_1 : \Omega \to \mathbb{R}$, $f_2 : \mathbb{R} \to \mathbb{R}$, $f_2|_{(-\infty,0)} = 0$, if both $f_1 : \Omega \to \mathbb{R}$ and several derivatives of $f_1 : \Omega \to \mathbb{R}$ satisfy Dirichlet boundary conditions on Ω, the smoother one requires that the periodic extension of $(f_1)_{odd} : \Omega_2 \to \mathbb{R}$ be, the faster $\langle f_1, \phi_j \rangle_\Omega$ — the jth Fourier sine series coefficient of f_1 — decays in j. Estimates on the decay of Fourier sine series coefficients have immediate implications for the

smoothness of solutions to initial boundary value heat problems. In particular, if the IBVP has zero initial conditions, then the jth Fourier sine series coefficient of the IBVP solution decays two orders of magnitude faster than the jth Fourier sine series coefficient of $f_1 : \Omega \to \mathbb{R}$. In such systems that have smooth $(f_1)_{odd}$, the high frequency eigenmodes are virtually insignificant to the IBVP solution.

Continuous system theory for the heat equation generalizes elegantly to the discrete realm. Indeed, in semi-discrete heat equations with spatially smooth discrete forcing terms, the high frequency eigenmodes are unimportant to the state variable. As the jth eigenvalue of the discrete Laplacian is on the order of j^2, it follows that modal truncation can improve the stable forward Euler time-step for integration of $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ by orders of magnitude without sacrificing accuracy in the state variable. Modal reduction is generally not cost-effective when applied to a sufficiently large full-order model. Hence one should consider a preliminary step that reduces the dimension prior to performing a modal filtration step.

In the setting of the heat equation, moment matching through ordinary Arnoldi successfully reduces the model dimension while maintaining accurate output, but high magnitude eigenvalues persist in the moment-matched system. Yet modal truncation is much less costly for the reduced-order moment-matched system and can be applied to remove those high frequency influences.

Expansion of the transfer function of the k-dimensional moment-matched system and its l-dimensional modally filtered counterpart in terms that decouple the

influences of individual eigenmodes unveils an intelligent choice for the mode of truncation based upon the smoothness of the system. In particular, the parameter ζ_j (2.49) measures the collective importance of the $k-j$ highest frequency modes in the moment-matched system. By identifying the mode of smallest index l such that $\zeta_{l+1} \leq \epsilon$ is below some specified tolerance, in an automated fashion one eliminates modes $\{l+1, \ldots, k\}$ while maintaining a desired level of accuracy in the approximate output. The quantity ζ_j decays faster with j for moment-matched systems in which the forcing term and the output coefficient row vector are smoother. In all of my experiments, if the full-order discrete forcing term is smooth, then so is the moment-matched counterpart. (This is intuitive, given that the low frequency Ritz vectors of the moment-matched system approximate the low frequency eigenvectors of the full-order model — those modes that contribute most prominently to the full-sized state variable — to high accuracy.) Hence choosing the truncation mode l according to $\zeta_{l+1} \leq \epsilon$ is based fundamentally upon the level of smoothness of the full-order semi-discrete system.

The dual-stage process works remarkably well in the numerical experiment of Chapter 2, for which the stable forward Euler time-step improves by nearly three orders of magnitude, and the time-step needed for order 10^{-3} relative accuracy in approximating the exact output is also significantly improved. Moreover, the full-order dimension is reduced by two orders of magnitude from 1024 to five.

One may argue that, upon completing moment matching, $\widehat{\Sigma}$ is adequately small

to justify numerical integration using the backward Euler scheme (3.17), and that the feasibility of this approach neutralizes the value of the dual-stage procedure because the backward Euler scheme has no stability restriction if $\sigma(\widehat{\mathbf{A}}) \subset (0, \infty)$. Yet the modal truncation step following moment matching is valuable not just as a method to increase the stable time-step, but also because it allows for additional dimension reduction, and it eliminates spurious information produced by Lanczos during the moment matching step. Consider the situation explained in Figure 2.9 for the numerical example of Sections 2.7.1 and 2.7.2. In that setting, moment matching produces $\widehat{\mathbf{A}}$ whose spectrum is comprised mostly of high frequency Ritz values, and this is in spite of the negligible relevance of such modes to the full-order model. Even more concerning is the fact that moment matching by Lanczos regularly introduces spurious high frequency modes. In light of such phenomena, modal truncation is valuable even when integrating via an implicit scheme, for it eliminates any peripheral or spurious high frequency information. Moreover, through its added layer of dimension reduction, the modal truncation step allows one to identify the essence of a system.

One may also object that, rather than measure the performance of the dual-stage procedure against that of the forward Euler scheme without model reduction and/or modal filtering, one should use an *exponential integrator* to perform a more rigorous comparison [13]. In particular, there exist several *A-stable* exponential integrators. As is true for the backward Euler scheme, such methods carry the advantage of having no time-step limit to ensure stability and can be applied at a reduced cost following

moment matching (relative to their cost for the full-order model) [13, p. 4].

Without question, in some regard, certain exponential schemes do provide a more rigorous standard for comparison than does the forward Euler scheme without modal filtering and/or moment matching. Yet the dual-stage procedure is designed specifically to improve upon explicit schemes, and hence it is natural to measure its impact by benchmarking against explicit schemes.

To adequately compare the performance of the explicit dual-stage procedure with the performances of both implicit and exponential integrators, one would need to perform a more exhaustive study than is contained in this thesis. Such a review could lead to further enhancements of the dual-stage procedure given the potential for more rigorous benchmarking using exponential and implicit schemes.

The Setting of Highly Nonnormal Convection-Diffusion Equations

For the two convection-diffusion experiments considered in Chapter 3, the dual-stage procedure has a mixed performance. Evidently, the positioning of ϵ-pseudospectra immediately adjacent to the imaginary axis for small values of ϵ renders moment matching by ordinary Arnoldi ineffective at reducing the dimension while also maintaining high relative accuracy. Shifting the spectrum away from the imaginary axis renders moment matching by ordinary Arnoldi more profitable. In that setting, modal truncation following moment matching substantially improves the *stable* time-step, but it only mildly improves the time-step required for *accurate* integration of the

moment-matched system. In spite of these mediocre gains in the time-step required for accuracy, modal truncation *is* highly effective in reducing the dimension of the moment-matched system while *maintaining* accuracy.

Based upon the mixed performance of the dual-stage procedure in these preliminary experiments, one ought to apply the dual-stage procedure with caution in the setting of such highly nonnormal convection-diffusion problems. In the setting of such problems, if the differential operator has an eigenvalue lying close to the origin, alternatives to the dual-stage procedure are advisable.

Generalizing the Theory for the Heat Equation to Sturm-Liouville Problems, and Additional Open Problems

The theory developed through this research for single-input, single-output systems should generalize usefully to multiple-input, multiple-output (MIMO), linear, time-invariant systems. This topic is a source of related interesting problems.

One may object that the theory presented for the heat equation applies only to a narrow class of problems. Yet it readily generalizes to a subset of the broader class of *Sturm-Liouville* problems. My brief exposition of such equations is based upon [14, p. 25-44].

Recall the function space $\mathcal{Q} := \{g \in C^2(\Omega) : g|_{\partial\Omega} = 0\}$ from (2.1). Fix $\beta := 1$ so that $\Omega := [0, 1]$. The *Sturm-Liouville* differential operator $\mathrm{M} : \mathcal{Q} \to \mathrm{L}^2(\Omega)$ is given by

$$\mathrm{M}w := \triangle w + q(x)w. \tag{4.1}$$

The Dirichlet *Sturm-Liouville IBVP* is of the form

$$\begin{array}{lcl} w_t & = & \mathrm{M}w + f \quad \text{on} \quad \Omega \times [0,T], \\ w|_{\Omega\times\{0\}} & = & w_0, \\ w|_{\partial\Omega\times[0,T]} & = & 0, \\ w|_{\Omega\times(-\infty,0)} & = & 0 \end{array} \tag{4.2}$$

for given $f : \Omega \times \mathbb{R} \to \mathbb{R}$, $f|_{\Omega\times(-\infty,0)} = 0$, $w_0 : \Omega \to \mathbb{R}$, and $q : \Omega \to \mathbb{R}$. Observe that the heat equation is one example of a Sturm-Liouville problem.

The following theorem, which I state without proof, shows the close relationship between the eigenmodes of the Sturm-Liouville and heat operators on $\mathcal{Q}$, even when $q \neq 0$ in the Sturm-Liouville problem.

*Theorem 4.1 (**Eigenvalues and Eigenfunctions of** M)*

For any $q \in \mathrm{L}^2(\Omega)$, the Sturm-Liouville operator $\mathrm{M} : \mathcal{Q} \to \mathrm{L}^2(\Omega)$ has eigenfunctions of the form

$$\psi_j(x) := \psi_j(x,q) := \phi_j(x) + \mathcal{O}(j^{-1}), \qquad j \in \mathbb{N}, \qquad j \to \infty, \tag{4.3}$$

where $\phi_j(x) := \sqrt{2}\sin(j\pi x)$ is the jth eigenfunction of the heat operator $\mathrm{L} : \mathcal{Q} \to \mathrm{L}^2(\Omega)$. Moreover, $\{\psi_j\}_{j\in\mathbb{N}}$ is an orthonormal basis for $\mathrm{L}^2(\Omega)$. The corresponding eigenvalues are given by

$$\alpha_j := \alpha_j(q) := \lambda_j + \int_\Omega q(x)dx + \xi_j, \tag{4.4}$$

where $\lambda_j := -(j\pi)^2$ is the jth eigenvalue of the heat operator, and $(\xi_1, \xi_2, \xi_l, \ldots) \in \ell^2$

is a sequence of real numbers such that

$$\sum_{j\in\mathbb{N}} \xi_j^2 < \infty.$$

See for example, [14, p. 35 and 43], for a more detailed explanation.

In particular, Theorem 4.1 shows that the jth eigenvalues and eigenfunctions of the Sturm-Liouville and heat operators increasingly resemble one another as $j \to +\infty$. The eigenvalues α_j and λ_j are said to be *asymptotically equivalent* as $j \to +\infty$ in the sense that $\lim_{j\to+\infty} \left(\alpha_j - \int_\Omega q(x)dx\right)/\lambda_j = 1$ [5, p. 10].

Whenever $q \in \mathrm{L}^2(\Omega)$, because $\{\psi_j\}_{j\in\mathbb{N}}$ forms an orthonormal basis for $\mathrm{L}^2(\Omega)$, if $w(\cdot,t) \in \mathrm{L}^2(\Omega)$ for all $t \geq 0$, one can solve the continuous Sturm-Liouville IBVP using a Fourier series expansion. In particular,

$$w(x,t) = \sum_{j\in\mathbb{N}} r_j(t)\psi_j(x),$$

where $r_j(t) := \langle w(\cdot,t), \psi_j\rangle_\Omega$ for $t \geq 0$ and $r_j(t) := 0$ for all $t < 0$.

Using the theory and ideas developed in Chapter 2, one can readily obtain estimates on the decay of $|r_j(t)|$ as $j \to +\infty$. For example, suppose that $w(x,0) = 0$, and consider the symbolic representation for the jth Fourier series coefficient given by

$$r_j(t) \;=\; \int_0^t e^{(t-s)\alpha_j}\widetilde{r}_j(s)ds$$

for $t \geq 0$, which can be verified by using an analogous approach to that used for (2.8). Here $\widetilde{r}_j(t) := \langle f(\cdot,t), \psi_j \rangle_\Omega$ for $t \geq 0$ and $\widetilde{r}_j(t) := 0$ for $t < 0$ is the jth coefficient in the Fourier series expansion

$$f(x,t) = \sum_{j \in \mathbb{N}} \widetilde{r}_j(t) \psi_j(x).$$

As in Chapter 2, define $\widetilde{s}_j(t) := \langle f(\cdot,t), \phi_j \rangle_\Omega$ for $t \geq 0$ and $\widetilde{s}_j(t) := 0$ for $t < 0$ to be the jth Fourier sine series coefficient for $f : \Omega \times \mathbb{R} \to \mathbb{R}$. A bit of analysis reveals that, assuming $\sup_{(x,t) \in \Omega \times \mathbb{R}} |f(x,t)| < +\infty$, as $j \to \infty$

$$r_j(t) = \mathcal{O}\left(j^{-2} \left\| \widetilde{s}_j(\cdot) \right\|_{\mathrm{L}^\infty(-\infty,t]} \right) + \mathcal{O}\left(j^{-3} \right). \tag{4.5}$$

Thus, by estimating the decay of the sine series coefficients for the forcing term $f : \Omega \times \mathbb{R} \to \mathbb{R}$ (as can be done through the results of Chapter 2), one generates a time-independent estimate on the decay of the coefficients $\{r_j(t)\}_{j \in \mathbb{N}}$.

The theory of Chapter 2 for the heat equation will readily generalize to linear, time-invariant systems generated from Sturm-Liouville IBVPs for which (4.5) predicts that the higher frequency modes are largely unimportant. Hence Sturm-Liouville IBVPs are a potentially rich source of further interesting numerical experiments involving the dual-stage reduction algorithm and of theoretical questions related to smoothness.

Appendix A

Appended Proofs

A.1 Proof Theorem 2.1

I proceed by induction using ideas from Gockenbach's related proof [10, p. 572-573]. Consider the base case, i.e., $f_{per} \in \mathrm{C}^0(\mathbb{R})$, and f'_{per} is piecewise continuous and bounded on $\mathbb{R}$. Using integration by parts and $\int \Phi_j = \frac{\beta}{-i\pi n}\Phi_j$, observe that

$$\begin{aligned} c_j &:= \left(\frac{1}{2\beta}\right)\left\langle f, \overline{\Phi_j}\right\rangle_{\Omega_2} \\ &= \left(\frac{1}{2\beta}\right)\left(\frac{\beta}{i\pi j}\right)\left(\int_{\Omega_2} f'(x)\Phi_j(x)dx - f(x)\Phi_j(x)\Big|_{\partial\Omega_2}\right). \end{aligned}$$

Now because $f_{per} : \mathbb{R} \to \mathbb{R}$ is continuous, and $\Phi_j(x)$ has period 2β,

$$\begin{aligned} f(x)\Phi_j(x)\Big|_{\partial\Omega_2} &= \Phi_j(\beta)\lim_{x\,\uparrow\,\beta} f(x) - \Phi_j(-\beta)\lim_{x\,\downarrow -\beta} f(x) \\ &= \Phi_j(\beta)f(\beta) - \Phi_j(\beta)f(\beta). \end{aligned}$$

Hence

$$c_j = \left(\frac{1}{2\beta}\right)\left(\frac{\beta}{i\pi j}\right)\left\langle f', \overline{\Phi_j}\right\rangle_{\Omega_2}.$$

Suppose that when $f_{per} : \mathbb{R} \to \mathbb{R}$ is $C^{p-3}(\mathbb{R})$ and $f_{per}^{(p-2)} : \mathbb{R} \to \mathbb{R}$ is piecewise

smooth and bounded on $\mathbb{R}$,

$$c_j = \left(\frac{1}{2\beta}\right)\left(\frac{\beta}{i\pi j}\right)^{p-2}\left\langle f^{(p-2)}, \overline{\Phi_j}\right\rangle_{\Omega_2}.$$

Then if $f_{per} : \mathbb{R} \to \mathbb{R}$ is $C^{p-2}(\mathbb{R})$, and $f_{per}^{(p-1)} : \mathbb{R} \to \mathbb{R}$ is piecewise smooth and bounded on $\mathbb{R}$, using integration by parts, observe that

$$c_j = \left(\frac{1}{2\beta}\right)\left(\frac{\beta}{i\pi j}\right)^{p-1}\left(\int_{\Omega_2} f^{(p-1)}(x)\Phi_j(x)dx - f^{(p-2)}(x)\Phi_j(x)\Big|_{\partial\Omega_2}\right).$$

Now $f^{(p-2)}(x)\Phi_j(x)\Big|_{\partial\Omega_2} = 0$ by the assumption that $f_{per}^{(p-2)}$ is continuous and using identical reasoning to that of the base case. Hence

$$c_j = \left(\frac{1}{2\beta}\right)\left(\frac{\beta}{i\pi j}\right)^{p-1}\left\langle f^{(p-1)}, \overline{\Phi_j}\right\rangle_{\Omega_2},$$

establishing the claim. ♦

A.2 Proof Theorem 2.2

I prove the two scenarios separately.

(i) Observe that, if $p \geq 3$, then $f : \Omega_2 \to \mathbb{R}$ satisfies the hypotheses of Theorem 2.1 for $p-1$. Namely, $f_{per} : \mathbb{R} \to \mathbb{R}$ is $\mathrm{C}^{p-3}(\mathbb{R})$ with $f_{per}^{(p-2)} : \mathbb{R} \to \mathbb{R}$ piecewise smooth

and bounded on $\mathbb{R}$ so that

$$c_j \;=\; \left(\frac{1}{2\beta}\right)\left(\frac{\beta}{i\pi j}\right)^{p-2}\left\langle f^{(p-2)},\overline{\Phi_j}\right\rangle_{\Omega_2}.$$

(The case of $p=2$ is identical, except one begins with the definition $c_j := \frac{1}{2\beta}\left\langle f,\overline{\Phi_j}\right\rangle_{\Omega_2}$ rather than use Theorem 2.1.) For fixed $\epsilon>0$, define the set

$$\Omega_2(\epsilon) := \Omega_2 \setminus \Big[[-\beta,-\beta+\epsilon) \bigcup \mathrm{B}_\epsilon(-a) \bigcup \mathrm{B}_\epsilon(0) \bigcup \mathrm{B}_\epsilon(a) \bigcup (\beta-\epsilon,\beta]\Big],$$

where for any $r\in\mathbb{R}$, $\mathrm{B}_\epsilon(r) := (r-\epsilon, r+\epsilon)$ is the ϵ-ball around r. Consider the term

$$\begin{aligned}\left\langle f^{(p-2)},\overline{\Phi_j}\right\rangle_{\Omega_2} &= \left\langle f^{(p-2)},\overline{\Phi_j}\right\rangle_{\Omega_2(\epsilon)} + \left\langle f^{(p-2)},\overline{\Phi_j}\right\rangle_{\Omega_2\setminus\Omega_2(\epsilon)}\\ &=: \mathrm{A}+\mathrm{B},\end{aligned}$$

where $\epsilon>0$ is arbitrary. The term $\mathrm{B}=\mathcal{O}(\epsilon)$ as $\epsilon\to 0$, as the integral of a bounded quantity over a domain of length 8ϵ.

To estimate the term A, observe that

$$\begin{aligned}\mathrm{A} &:= \int_{\Omega_2(\epsilon)} f^{(p-2)}(x)\Phi_j(x)dx\\ &= \frac{\beta}{i\pi n}\left(\int_{\Omega_2(\epsilon)} f^{(p-1)}(x)\Phi_j(x)dx \;-\; f^{(p-2)}(x)\Phi_j(x)\Big|_{\partial(\Omega_2(\epsilon))}\right)\\ &=: \mathrm{C}+\mathrm{D},\end{aligned}$$

which follows by using integration by parts and the observation that $\int\Phi_j = \frac{\beta}{-i\pi j}\Phi_j$.

Now

$$|\mathrm{C}| \;\leq\; \frac{2\beta^2}{\pi|j|}\|f^{(p-1)}\|_{\mathrm{L}^\infty(\Omega_2(\epsilon))},$$

and

$$\mathrm{D} \;=\; -\frac{\beta}{i\pi j}\left(\begin{array}{c} f^{(p-2)}(x)\Phi_j(x)\Big|_{-\beta}^{-\beta+\epsilon} + f^{(p-2)}(x)\Phi_j(x)\Big|_{-a-\epsilon}^{-a+\epsilon} \\ +f^{(p-2)}(x)\Phi_j(x)\Big|_{-\epsilon}^{\epsilon} + f^{(p-2)}(x)\Phi_j(x)\Big|_{a-\epsilon}^{a+\epsilon} \\ +f^{(p-2)}(x)\Phi_j(x)\Big|_{\beta-\epsilon}^{\beta} \end{array}\right).$$

Observe that the derivative of an odd (even) function is itself even (odd). It then follows that $f_{per}^{(p-2)} : \mathbb{R} \to \mathbb{R}$ is either even or odd, implying that $|f^{(p-2)}(x)| = |f^{(p-2)}(-x)|$ for all x. Coupling this observation with the fact that $\|\Phi_j\|_{\mathrm{L}^\infty(\Omega_2)} = 1$,

$$|\mathrm{D}| \;\leq\; \frac{2\beta}{\pi|j|}\left(\begin{array}{l} |f^{(p-2)}(\beta)| + |f^{(p-2)}(\beta-\epsilon)| + |f^{(p-2)}(a+\epsilon)| \\ +|f^{(p-2)}(a-\epsilon)| + |f^{(p-2)}(\epsilon)| \end{array}\right).$$

Hence for arbitrary $\epsilon > 0$,

$$c_j \;=\; \left(\frac{1}{2\beta}\right)\left(\frac{\beta}{i\pi j}\right)^{p-2}[\mathrm{C}+\mathrm{D}+\mathrm{B}],$$

and

$$|c_j| \;\leq\; \left(\frac{1}{2\beta}\right)\left(\frac{\beta}{\pi|j|}\right)^{p-2}\left[\begin{array}{l} \frac{2\beta^2}{\pi|j|}\|f^{(p-1)}\|_{\mathrm{L}^\infty(\Omega_2(\epsilon))} \\ \\ +\frac{2\beta}{\pi|j|}\left(\begin{array}{l} |f^{(p-2)}(\beta)|+|f^{(p-2)}(\beta-\epsilon)|+|f^{(p-2)}(a+\epsilon)| \\ \\ +|f^{(p-2)}(a-\epsilon)|+|f^{(p-2)}(\epsilon)| \end{array}\right) \\ \\ +|\mathrm{B}| \end{array}\right],$$

so that

$$|c_j| \;\leq\; \frac{\beta^{p-2}}{(\pi|j|)^{p-1}}\left(\begin{array}{l} \beta\|f^{(p-1)}\|_{\mathrm{L}^\infty(\Omega_2\setminus\{0,\ \pm a,\ \pm\beta\})} \\ \\ +|f^{(p-2)}(0)|+2|f^{(p-2)}(a)|+2|f^{(p-2)}(\beta)| \end{array}\right),$$

where I have used the assumption that $f_{per}^{(p-2)} \in C^0(\mathbb{R})$ and taken the limit as $\epsilon \to 0$.

(ii) Theorem 2.1 implies that

$$c_j \;=\; \left(\frac{1}{2\beta}\right)\left(\frac{\beta}{i\pi j}\right)^{p-1}\left\langle f^{(p-1)}, \overline{\Phi_j}\right\rangle_{\Omega_2}.$$

Analogously to case **(i)**, consider the expansion

$$\begin{aligned} \left\langle f^{(p-1)}, \overline{\Phi_j}\right\rangle_{\Omega_2} &= \left\langle f^{(p-1)}, \overline{\Phi_j}\right\rangle_{\Omega_2(\epsilon)} + \left\langle f^{(p-1)}, \overline{\Phi_j}\right\rangle_{\Omega_2\setminus\Omega_2(\epsilon)} \\ &=: \mathrm{A}+\mathrm{B}, \end{aligned}$$

where $\epsilon > 0$ is arbitrary. By identical reasoning to that of case (i), $\lim_{\epsilon\to 0}\mathrm{B} = 0$.

Recalling that all potential discontinuities of $f^{(p-1)}$ and $f^{(p)}$ on Ω_2 are isolated to

the set of singleton the points $\{0,\ \pm a,\ \pm\beta\}$, one can integrate by parts to find that

$$\begin{aligned} \mathrm{A} &= \frac{\beta}{i\pi j}\left(\int_{\Omega_2(\epsilon)} f^{(p)}(x)\Phi_j(x)dx \ - \ f^{(p-1)}(x)\Phi_j(x)\Big|_{\partial(\Omega_2(\epsilon))}\right) \\ &=: \ \mathrm{C}+\mathrm{D}, \end{aligned}$$

where, as in case **(i)**,

$$|\mathrm{C}| \ \leq \ \frac{2\beta^2}{\pi|j|}\|f^{(p)}\|_{\mathrm{L}^\infty(\Omega_2(\epsilon))}.$$

Now

$$\mathrm{D} \ = \ -\frac{\beta}{i\pi j}\left(\begin{array}{l} f^{(p-1)}(x)\Phi_j(x)\Big|_{-\beta}^{-\beta+\epsilon} + f^{(p-1)}(x)\Phi_j(x)\Big|_{-a-\epsilon}^{-a+\epsilon} + f^{(p-1)}(x)\Phi_j(x)\Big|_{-\epsilon}^{\epsilon} \\ \\ +f^{(p-1)}(x)\Phi_j(x)\Big|_{a-\epsilon}^{a+\epsilon} + f^{(p-1)}(x)\Phi_j(x)\Big|_{\beta-\epsilon}^{\beta} \end{array}\right),$$

and

$$|\mathrm{D}| \ \leq \ \frac{10\beta\|f^{(p-1)}\|_{\mathrm{L}^\infty(\mathbb{R})}}{\pi|j|}.$$

Thus

$$c_j \ = \ \left(\frac{1}{2\beta}\right)\left(\frac{\beta}{i\pi j}\right)^{p-1}\left[\mathrm{C}+\mathrm{D}+\mathrm{B}\right],$$

and taking the limit as $\epsilon \to 0$,

$$|c_j| \leq \frac{\beta^{p-1}}{(\pi|j|)^p} \left(\begin{array}{l} \beta \|f^{(p)}\|_{\mathrm{L}^\infty(\Omega_2 \setminus \{0,\ \pm a,\ \pm\beta)} \\ \\ +5\|f^{(p-1)}\|_{\mathrm{L}^\infty(\mathbb{R})} \end{array} \right).$$

♦

A.3 Proof of Theorem 3.1

Throughout this proof, I adopt the notation that Gockenbach uses for second-order linear homogeneous ODEs [10, p. 85-87].

Observe that $\mathrm{H}\phi = \lambda\phi$ with $\phi \neq 0$ is true if and only if

$$d\frac{\mathrm{d}^2}{\mathrm{d}x^2}\phi + c\frac{\mathrm{d}}{\mathrm{d}x}\phi - \lambda\phi = 0, \tag{A.1}$$

an ODE with characteristic roots [10, p. 85]

$$\begin{aligned} r_1 &:= \frac{-c - \sqrt{c^2 + 4d\lambda}}{2d}, \\ r_2 &:= \frac{-c + \sqrt{c^2 + 4d\lambda}}{2d}. \end{aligned}$$

Now if $c^2 + 4d\lambda > 0$, then all solutions to (A.1) are of the form

$$\phi(x) = c_1 e^{r_1 x} + c_2 e^{r_2 x},$$

where c_1, $c_2 \in \mathbb{C}$ [10, case one on p. 86]. Observe that $\phi(0) = 0$ gives $-c_1 = c_2$. The

second boundary condition,

$$0 = \phi(\beta) = c_1 \left(e^{\beta r_1} - e^{\beta r_2} \right),$$

together with $\phi \neq 0$, implies that $r_1 = r_2$. Namely, $c^2 + 4d\lambda = 0$, which is a contradiction. Thus $c^2 + 4d\lambda \leq 0$.

If $c^2 + 4d\lambda = 0$, then all solutions to (A.1) are of the form

$$\phi(x) \quad = \quad c_1 e^{rx} + c_2 x e^{rx},$$

where $r = r_1 = r_2$ and $c_1,\ c_2 \in \mathbb{C}$ [10, case three on p. 86-87]. The boundary conditions imply that $0 = \phi(0) = c_1$, and $0 = \phi(\beta) = c_2 e^r$. Yet $e^r \neq 0$ implies $c_2 = 0$, which contradicts $\phi \neq 0$.

It follows that $c^2 + 4d\lambda < 0$. In this scenario, the general solution to (A.1) is given by

$$\phi(x) \quad = \quad c_1 e^{\mu_1 x} \cos(\mu_2 x) + c_2 e^{\mu_1 x} \sin(\mu_2 x),$$

where $c_1,\ c_2 \in \mathbb{C}$, and $r_1 = \mu_1 - i\mu_2$ and $r_2 = \mu_1 + i\mu_2$ for

$$\begin{aligned} \mu_1 &:= -\frac{c}{2d} \in \mathbb{R}, \\ \mu_2 &:= \frac{1}{i} \frac{\sqrt{c^2 + 4d\lambda}}{2d} \in \mathbb{R} \end{aligned}$$

[10, case two on p. 86].

Observe that $0 = \phi(0) = c_1$. Moreover, $0 = \phi(\beta) = c_2 e^{\beta\mu_1} \sin(\beta\mu_2)$, together with $\phi \neq 0$ and $e^{\beta\mu_1} \neq 0$, implies that $\sin(\beta\mu_2) = 0$, which can be true if and only if

$$\mu_2 \in \left\{ \frac{z\pi}{\beta} \right\}_{z\in\mathbb{Z},\ z\neq 0}.$$

Denote $\mu_2(z) := \frac{z\pi}{\beta}$ and observe that for all nonzero $z \in \mathbb{Z}$, $\phi_z := e^{\mu_1 x} \sin(\mu_2(z)x)$ is an eigenfunction of H with corresponding eigenvalue λ_z defined by the equation

$$\frac{1}{i}\frac{\sqrt{c^2 + 4d\lambda_z}}{2d} \quad = \quad \mu_2(z).$$

Immediately,

$$\lambda_z \quad = \quad -\left(\frac{(2d\pi z/\beta)^2 + c^2}{4d} \right).$$

Note that $\lambda_{-z} = \lambda_z$ and $\phi_{-z} = -\phi_z$, so that z and $-z$ correspond to identical eigenpairs.

For the final claim, suppose that for some set of scalars, $\{c_j\}_{j\in\mathbb{N}} \subset \mathbb{C}$,

$$0 = \sum_{j\in\mathbb{N}} c_j \phi_j(x) = e^{-\frac{cx}{2d}} \sum_{j\in\mathbb{N}} c_j \sin(\pi j x/\beta).$$

Then $e^{-\frac{cx}{2d}} \neq 0$ implies that $\sum_{j\in\mathbb{N}} c_j \sin(\pi j x/\beta) = 0$. By the fact that $\{\sin(\pi j x/\beta)\}_{j\in\mathbb{N}}$ forms a set of orthogonal functions on Ω (see discussion surrounding (2.3)), $c_j = 0$

for all j. The claim follows. ♦

Bibliography

[1] Antoulas, Athanasios C. *Approximation of Large-Scale Dynamical Systems*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2005.

[2] Bonvin, D., and D.A. Mellichamp. "A Unified Derivation and Critical Review of Modal Approaches to Model Reduction." *International Journal of Control* 35, no. 5 (1982): 829-848.

[3] Bracewell, Ronald N. *The Fourier Transform and Its Applications.* 2nd ed., Revised. Boston, MA: WCB/McGraw-Hill, 1978.

[4] Briggs, William L., and Van Emden Henson. *The DFT: An Owner's Manual for the Discrete Fourier Transform*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 1995.

[5] De Bruijn, N.G. *Asymptotic Methods in Analysis*. 3rd ed., unabridged and corrected. New York, NY: Dover Publications, Inc., 1981.

[6] Embree, Mark. "Lecture Notes: CAAM 453: Numerical Analysis I." Rice University, Fall 2009.

[7] Embree, Mark, and D.C. Sorensen. *An Introduction to Model Reduction for Linear and Nonlinear Differential Equations*. To appear, draft from spring 2011.

[8] Embree, M., and D.C. Sorensen. "Model Reduction Experiments: Part II: Moment Matching Reduction," *CAAM 651: Topics in Numerical Linear Algebra: Numerical Methods for Dimension Reduction of Dynamical Systems*, Department of Computational and Applied Mathematics, Rice University, Spring 2011, `http://www.caam.rice.edu/~caam651/modred_tasks2.tex` (accessed July 27, 2012).

[9] Gallopoulos, E., and Y. Saad. "Efficient Solution of Parabolic Equations by Krylov Approximation Methods." *SIAM J. Sci. and Stat. Comput.* 13, no. 5 (1992): 1236-1264.

[10] Gockenbach, Mark S. *Partial Differential Equations: Analytical and Numerical Methods*. 2nd ed. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2011.

[11] Green, Michael, and David J. N. Limebeer. *Linear Robust Control*. Englewood Cliffs, NJ: Prentice Hall, 1995.

[12] LeVeque, Randall J. *Finite Difference Methods for Ordinary and Partial Differential Equations: Steady-State and Time-Dependent Problems*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2007.

[13] Minchev, Borislav V., and Will M. Wright. "A Review of Exponential Integrators for First Order Semi-Linear Problems." TR No. 2/2005 (Preprint), Department of Mathematical Sciences, Norwegian University of Science and Technology, 2005. `http://www.math.ntnu.no/preprint/numerics/2005/N2-2005.ps` (accessed October 3, 2012).

[14] Pöschel, Jürgen, and Eugene Trubowitz. *Inverse Spectral Theory*. Orlando, FL: Academic Press, Inc., 1987.

[15] Reddy, Satish C., Lloyd N. Trefethen, and Dimpy Pathria. "Pseudospectra of the Convection-Diffusion Operator." TR 93-1337, Department of Computer Science, Cornell University, 1993. `http://hdl.handle.net/1813/6103` (accessed November 20, 2012).

[16] Royden, H.L. *Real Analysis*. 3rd ed. Englewood Cliffs, NJ: Prentice Hall, 1968.

[17] Süli, Endre, and David Mayers. *An Introduction to Numerical Analysis*. Cambridge: Cambridge University Press, 2006.

[18] Weisstein, Eric W. "Factorial Sums." From `MathWorld`—A Wolfram Web Resource. `http://mathworld.wolfram.com/FactorialSums.html` (accessed March 8, 2012).

[19] Trefethen, Lloyd N., and Mark Embree. *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*. Princeton, NJ: Princeton University Press, 2005.

[20] Trefethen, Lloyd N. *Spectral Methods in Matlab*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2000.

[21] Trefethen, Lloyd N., and David Bau, III. *Numerical Linear Algebra*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 1997.